Mental Illness and Suicide Ideation Detection Using Social Media Data

by

Prasadith Buddhitha Kirinde Gamaarachchige

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for the
Doctor of Philosophy in Digital Transformation and Innovation

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Abstract

Mental disorders and suicide have become a global public health problem. Over the years, researchers in computational linguistics have extracted features from social media data for the early detection of users susceptible to mental disorders and suicide ideation. Lack of reliable and inadequate data and the requirement of interpretability can be identified as the principal reasons for the low adoption of neural network architectures in recognizing individuals with mental disorders and suicide ideation. In recent years, a gradual increase in the use of deep neural network architectures in detecting mental disorders and suicide ideation with low false positive and false negative rates became feasible. Our research investigates the efficacy of using a shared representation to learn lower-level features mutual among mental disorders and between mental disorders and suicide ideation. In addition to discovering the shared features between users with suicidal thoughts and users who self-declared a single mental disorder, we further investigate the impact of comorbidities on suicide ideation and use two unseen datasets to investigate the generalizability of the trained models. We use data from two different social media platforms to identify if knowledge can be shared between suicide ideation and mental illness detection tasks across platforms. Through multiple experiments with different but related tasks, we demonstrate the effectiveness of multi-task learning (MTL) when predicting users with mental disorders and suicide

ideation. We produce competitive results using MTL with hard parameter sharing when predicting neurotypical users, users who might have PTSD (Post-Traumatic Stress Disorder), and users with depression. The results were further improved by using auxiliary inputs such as emotion, age, and gender. To predict users with suicide ideation or mental disorders (i.e., either single or multiple disorders), we use MTL with hard and soft parameter sharing and produce state-of-the-art results predicting users with suicide ideation who require urgent attention. For similar tasks, but with data from two different social media platforms, we further improve the state-of-the-art results when predicting users with suicide ideation who require urgent attention. In addition, we managed to improve the overall performances of the models by using different auxiliary inputs.

Acknowledgments

I would like to dedicate my thesis to my supervisor Prof. Diana Inkpen, for introducing me to this amazing field of Natural Language Processing and opening my eyes to the area of Artificial Intelligence. Since 2014 when Prof. Diana accepted me as one of her master's students, and till this day, I have received immense support, knowledge, advice, and guidance. Because of her guidance and advice, I participated in conferences, workshops, and shared tasks from which I obtained invaluable knowledge. Writing this thesis will not be possible unless Prof. Diana has introduced me to the research area of mental illness and suicide ideation detection, where we can make valuable contributions to the research community and, more importantly, to society. Also, I am profoundly thankful to Prof. Diana for providing me with research funding throughout the years.

I would sincerely like to thank Prof. Liam Peyton for all the guidance, advice, and support he has provided throughout my journey at the University of Ottawa. From the courses he taught to the invaluable advice has helped me to complete my research work.

I would like to thank my thesis committee members, Prof. Morad Benyoucef, Prof. Hussein Al Osman and Prof. Marie-Jean Meurs, for the insightful comments, feedback, and guidance that helped me produce a strong thesis.

I will not be able to conduct my research without the amazing help from Dr. Jelber Sayyad Shirabad, who resolved all the infrastructure issues that I faced when running the deep learning models. Also, I am thankful for the insightful discussions we had over the years that helped me grow as a better researcher.

Great thanks go to Ahmed Husseini Orabi and Mahmoud Husseini Orabi for their tremendous help and support throughout the years as collaborators and colleagues. With many conversations we had throughout the years, I gained more knowledge about the field of artificial intelligence and, specifically, deep learning.

This journey as a researcher would not be possible if not for the great sacrifices made by my wife, Thakshayini. I am ever grateful to my wife for encouraging me to pursue my journey to become a researcher and supporting me exceedingly throughout these years. I am incredibly thankful to my kids Thevinshya, Thevinya and Partheeshan, for their understanding of the time and energy I had to contribute throughout the years.

I am ever grateful to my mother Turin, my father Piyasena and my sister Chaminie for encouraging my journey in the pursuit of knowledge and my father-in-law Karunagaran and my mother-in-law Nagulambigai for their continuous support.

Great thanks to my labmate Ehsan Amjadian for being an amazing friend throughout the years. Through our insightful discussions over the years, I gained a wealth of knowledge about NLP and machine learning. I would like to give my heartful gratitude to all my lab mates and colleagues for their support throughout the years. I would like to thank Zunaira Jamil, Ruba Skaik, Haifa Alharthi, Rex Liu, Arya Rahgozar, Vaibhav Kesarwani, Saman Daneshvar, Shy Huang, Ken Xin and Cheng Duan, and all the TAMALE group members.

Finally, I am thankful to the University of Ottawa for grating me with the

International Doctoral Scholarship and providing a safe environment to pursue my journey as a researcher.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

1.1.1 Suicide Ideation Detection

Suicide and mental health are formidable challenges faced by the whole world. In 2019, 1.3% of all deaths were due to suicide, which was one of the leading causes of death worldwide. The suicide rate varies from country to country. It is around two deaths per 100,000 persons for some countries, while others have reported around 80 deaths per 100,000 persons. The reported figures are for all the gender and age groups. Globally, more suicides have been reported from countries with low to middle income, and when considering the age groups, more than 50% of the suicides were committed by individuals before they were 50 years old (World Health Organization, 2021).

In Canada, there were 4,012 suicides reported in 2019 and out of which 3,058 were male. Overall, for every 100,000 persons, 10.7 suicides were reported, and according to gender, it was 16.4 males and 5.0 females. Of 100,000, more than 15 individuals have committed suicide in the age groups 50-54 and 55-59. However, on average, around 13 individuals for every 100,000 within the age groups from 20 to 49 have

committed suicide. To further emphasize the severity of the suicide problem, the Canadian community health survey has reported that 1 in every 10 Canadians above 15 years of age has thought about suicide during their lifetime. Similarly, around 27% of the First Nations people living in reserve areas and above 15 years of age have thought about suicide in their lifetime (Statistics Canada, 2020). The statistics mentioned above and more recent series of tragic events where five students at the University of Ottawa have taken their lives within less than a year (i.e., from the beginning of 2019) (Dubé, 2020; Yogaretnam, 2020), and more recently, the death of the sixth student for which the cause of death has not been confirmed to the public (Dubé, 2020), highlight the relevance of the early detection of suicidal behaviour.

The report of the World Health Organization (2021) introduced four mediations to prevent suicide, and out of which, early identification of suicidal behaviour, assessment, management, and follow-up can be collectively considered as one of the critical points. One of the main objectives in our research is to predict users with suicide ideation, and as a result, we will be able to use our trained models to identify users at risk of suicide as a preliminary step in suicide prevention.

1.1.2 Mental Illness Detection

When evaluating the suicide risk factors, it has been identified that mental disorders are strongly correlated with suicide attempts. Even though adequate research has not been conducted to rank mental disorders based on their impact on suicide attempts, Nock et al. (2009) identified that being diagnosed with a mental illness increases the risk of suicidal behaviour. Extensive research has been conducted to identify the level of impact that mental disorders have on suicidal behaviour where research on

post-traumatic stress disorder (PTSD) (LeBouthillier et al., 2015), bipolar (Dome et al., 2019), depression (Hawton et al., 2013) and schizophrenia (Zaheer et al., 2020) have shown that mental disorders are indeed associated with increased levels of suicidality. Further research has been conducted to identify the impact comorbid conditions have on suicide, and the research outcomes have shown that comorbidity of disorders or having more than one disorder at a given time increases the risk of suicide (Brådvik, 2018; Holmstrand et al., 2015). To identify and evaluate the before-mentioned relationship between mental disorders and suicide ideation, we conducted several experiments using data from users who self-reported diagnoses of single and multiple mental disorders. Given the impact mental disorders have on suicide ideation, it could be argued that detecting users with mental illnesses is as essential as predicting users with suicide risk, where early detection and treatment of users with mental disorders could reduce the severe impact it might have on their mental and physical wellbeing.

In general, the mental health of Canadians has been declining: when comparing the statistics between 2015 and 2019, around 5% of a decline in mental health is identified among people aged 12 and above. When analyzing the type of mental disorders Canadians aged above 12 are diagnosed with, in the year 2019, around 14% have reported being diagnosed with mood and anxiety disorders. Compared to 2015, it is an increase of 2%, and the surge was more significant among individuals aged between 18 and 34, which is around 4% (Statistics Canada, 2020). Because early detection and treatment are critical prevention mechanisms to lower the impact mental illnesses have on society (World Health Organization, 2004), we introduce our research in mental illness detection to early detect users susceptible to mental disorders.

Analyzing the age groups, we could see that many people aged 20-49 have suicidal thoughts or are diagnosed with a mental disorder. If diagnosed, a person needs to receive the necessary care. According to Statistics Canada (2019), in 2018, out of 5.3 million people who have indicated that they require help to overcome their mental health issues, 1.1 million people have reported that they could not get the necessary support. In addition, around 1.2 million Canadians have indicated that even though they received a certain level of support, it was insufficient to resolve their mental health issues fully. Some of the reasons for not receiving the necessary support are the associated cost of treatment and the lack of information on how to obtain the necessary help. In addition, social stigma and discrimination have also prohibited people from getting the required treatments and social support (World Health Organization, 2018). Given these circumstances, we think it is essential to find an effective solution to identify people who require help and provide them with the relevant information and guidance as a preliminary step into their path of recovery.

1.1.3 Impact of Social Media

Social media platforms have revolutionized the way people interact as a society and have become an integral part of the everyday life of many. People have started sharing their day-to-day activities on these platforms, which as a result, could reveal invaluable insights into one's cognition, emotion, and behavioural aspects. With its rapid growth among different demographics and being a source enriched with valuable information, social media can be a significant contributor to the process of mental disorder and suicide ideation detection. When analyzing the social media usage within the Canadian population, for every 10 individuals, 9 and 8 in the age groups 15 to

34 and 35 to 49, respectively, are using social media platforms (Schimmele et al., 2021). Even though many are using various social media platforms, it is essential to identify to what extent they share information so that researchers can extract the relevant content that represents suicidal thoughts and mental illnesses. According to the findings of Schimmele et al. (2021), over 25% of social media users within the age group 15 to 64 publicly share personal content and especially around 33% of users in the age group 15 to 19 share more personal content than the rest of the social media users.

According to the before-mentioned statistical information, we identified the importance of early detecting users with mental disorders and suicide ideation to initiate the necessary prevention mechanisms. To detect at-risk users, we decided to use the social media platforms given their frequent use by individuals under different age groups and the extent of personal content shared on these platforms. The shared real-time content portraying one's daily life could reveal invaluable insights that could be difficult to obtain using other forms of data collection strategies such as structured questionnaires[1]. The questionnaires are often used as an initial step to screen individuals for mental illnesses. In such surveys, the interviewee could be more vulnerable to memory bias and adapt to the guidelines prescribed by the assessor. Using such screening procedures might not expose the actual mental state of an individual, and hence, the prescribed treatments could be inadequate.

[1]Both CLPSych 2015 and SMHD datasets contain self-reported mental health diagnoses, so that the diagnosis could have included answering the structured questionnaires.

1.1.4 Use of Deep Learning

In a domain with limited annotated data on mental disorders and suicide ideation and with the requirement of reasoning, researchers were restricted to using traditional machine learning approaches other than using more recent methods such as deep neural networks. Given the social requirement and the current use of machine learning technologies in the mental illness and suicide ideation detection domain, we explore the feasibility of applying deep learning methods and specifically multi-task learning to predict users with single or multiple mental disorders and suicide ideation. In addition, we explore the impact auxiliary inputs, specifically the ones discovered by researchers, have on mental illness and suicide ideation detection outcomes. We opted for multi-task learning, given the nature of the tasks where certain mental disorders share common characteristics. Also, individuals with suicidal thoughts have an increased chance of being diagnosed with single or multiple mental disorders (comorbidity).

Detecting and treating mental disorders and individuals with suicide ideation can be identified as a complex clinical decision. Considering the complexities and skills required, predicting mental illnesses and suicide ideation among individuals using natural language processing and machine learning techniques could be considered a preliminary step in generating awareness rather than deriving conclusions on one's mental state.

The constraints mentioned above and the research opportunities motivated us to use deep learning methods to detect mental disorders and suicide ideation using social media platforms (especially text data). We will highlight the opportunities presented when using deep neural networks and, specifically, different multi-task

learning architectures. We will also discuss the prospects of using auxiliary inputs discovered through exploratory analysis and manual feature engineering.

Even though it will be beneficial to have sufficient justification on the predictive outcome (which can be possible when using some traditional machine learning methods), it could be argued that in specific applications such as online support forums, accurately predicting individuals at high risk is more important than providing a less accurate prediction with explanations on the features deriving the outcome.

1.2 Problem Statement

Overall, we could see that the mental health support required by many Canadians is either unmet or partially met. Unless treated accordingly, the mental and physical health of the person diagnosed could further deteriorate. In order to provide the necessary support, it is essential to detect the individuals in need of help early. The widespread use of social media platforms and users sharing their personal content on these platforms have shaped the opportunity to recognize the users needing mental health support.

Given these opportunities, we conducted several experiments to detect users with mental disorders and suicide ideation. We trained a deep learning model to predict users who self-reported diagnoses of mental disorders using limited Twitter data. The predictions were to detect users of three types: with PTSD (post-traumatic stress disorder), with depression, or neurotypical users who are not diagnosed with either of the illnesses. To improve the prediction capabilities of the trained model, we used emotion categories predicted for each user.

With the necessity of early detecting users with suicide ideation, we used a dataset

extracted from the Reddit social media platform and annotated by crowdsourced and domain experts. The dataset was used to identify if a user had suicide ideation or not and if users with suicide ideation require urgent attention. We did not use this dataset by itself but combined it with data containing users who self-declared mental disorders. The self-declaration could be for single or multiple mental disorders, which presented us with further research opportunities to identify the impact of single and multiple mental disorders on suicide ideation. Assuming that when diagnosed with multiple mental disorders, the person can be diagnosed with two or more disorders at a given time provides the necessary environment to identify the impact comorbidity of disorders has on suicide ideation. We focused not only on identifying the impact mental disorders have on suicide ideation but also the impact suicidal thoughts have on mental disorders.

Finally, we conducted several experiments to explore the possibility of knowledge transfer between two different social media platforms with different distributions to identify the impact different mental disorders have on suicidal thoughts and the impact suicidal thoughts have on mental disorders. These experiments confirm that the social media platform is not relevant when discovering the bilateral impact between suicide ideation and mental disorders.

Rather than using classical machine learning methods with manually engineered features, we explored the possibility of using deep learning architectures and, specifically, multi-task learning, given its appropriateness in situations where each task could benefit from the others by sharing lower-level features.

Throughout our research, we address the requirement of early detecting users

diagnosed with mental disorders and suicide ideation to initiate the necessary assessment and prevention mechanisms. Implementing the proposed architecture in a safe platform where many can reach for help and support can be used as the first step to prevent suicide and provide help and guidance for the users diagnosed with single or multiple mental disorders. It is important to note that our proposed architecture should not be considered a prevention mechanism but only a solution to generate awareness so that the relevant authorities with trained personnel could provide the necessary support.

1.3 Research Questions

Given the before-mentioned problems that we identified by analyzing the current social challenges and their potential solutions, we will conduct our research to answer the following questions:

- Can multi-task learning with automatically-calculated auxiliary inputs be used effectively to predict users with different mental health conditions?
- Can multi-task learning with two datasets with different distributions be used to identify the impact mental disorders (i.e., single or multiple) have on suicide ideation detection, and the impact suicide ideation has on mental disorders detection?
- Can multi-task learning be used to transfer knowledge between two different social media platforms to discover the impact mental disorders have on suicide ideation detection and the impact suicide ideation has on mental disorders detection?

1.4 Research Overview

Figure 1.1 summarizes the conducted research that focuses on mental illness and suicide ideation detection. Our research used different datasets, as mentioned in figure 1.1. We used four datasets discussed in detail in section 3.1. The WASSA-2017 (Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis) dataset (Mohammad & Bravo-Marquez, 2017) was used to train a model that predicts the emotion categories of the individual tweets in the CLPSych 2015 (Computational Linguistics and Clinical Psychology) dataset (Coppersmith, Dredze, Harman, Kristy, & Mitchell, 2015). As one of our research objectives, we wanted to identify the mental disorder given Twitter social media data. In order to identify if a user is neurotypical or one who self-reported diagnoses (hereafter referred to as "diagnosed")[2] of either PTSD or depression, we used the CLPSych 2015 dataset in combination with emotion categories as auxiliary inputs. In addition, the CLPSych 2015 dataset/task was used as one of the tasks in a multi-task learning environment to identify users with either suicide ideation or mental disorders. Different combinations of data samples containing users diagnosed with PTSD or depression were used to identify the impact mental disorders have on suicide ideation. For the second task, we used the UMD dataset (i.e., The University of Maryland Reddit suicidality dataset) (Shing et al., 2018) that was divided into two parts based on the assigned risk level. We conducted several experiments by grouping data based on the suicide risk label, where the experiments for the flagged/not flagged tasks were implemented to identify users with and without suicide ideation. The urgent/not urgent task was to identify users with suicide ideation that require urgent attention. The dataset was annotated by

[2]We use the term diagnosed to identify users who have self-declared diagnoses in both CLPSych 2015 and SMHD datasets.

crowdsourced annotators and expert annotators. We used the crowdsource annotated data to train, validate, and test our models. Then, we used the expert annotated data as an additional test set to investigate further the models' generalizability (without re-training them). Using datasets from two different social media platforms with different distributions enabled us to investigate if knowledge can be shared successfully between different social media platforms.

To identify the impact mental disorders have on suicide ideation and also the impact suicide ideation has on mental disorders, we used the SMHD (Self-Reported Mental Health Diagnoses) dataset (Cohan et al., 2018). We further divided our experiments using the SMHD data to investigate the impact of single and multiple mental disorders on suicide ideation. Multiple mental disorders are where a user has self-declared more than one disorder through their published posts. Even though we could not find adequate evidence that the user has been diagnosed with all the disorders at a given time, we assume that a given user is diagnosed with at least two or more disorders at a given time, hence the impact of comorbidity of disorders is investigated.

1.5 Ethical Considerations

It is essential to follow strict guidelines on ethical research conduct, especially when the research data is about vulnerable users, including users diagnosed with mental disorders or suicidal thoughts. Researchers working with data that could single out individuals must take adequate precautions to avoid further psychological distress on those needing mental health support. Some researchers have taken sufficient steps in anonymizing the data to secure user privacy. Coppersmith, Dredze, Harman, Kristy,

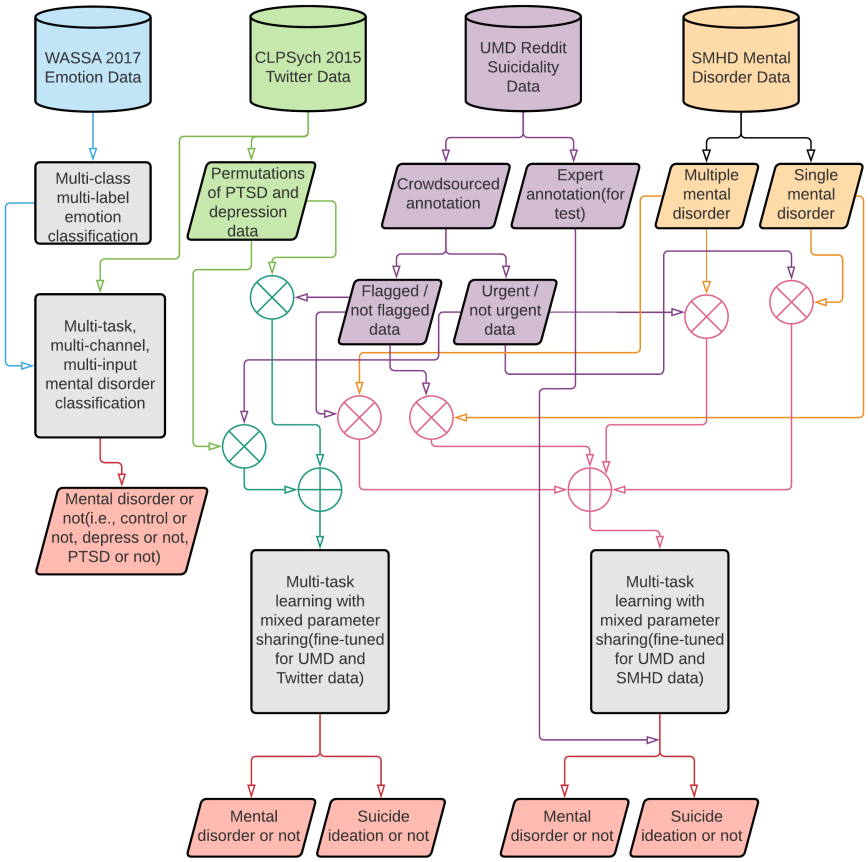


Figure 1.1: Research overview

and Mitchell (2015), have used a whitelist approach in anonymizing the data given to the CLPsych 2015 shared task participants. Even though screen names and URLs were anonymized using salted hash functions, the possibility of cross-referencing the hashed text against the Twitter archives still exists, and it could lead to a breach of user privacy. Due to this reason, the shared task participants were asked to sign a confidentiality agreement to ensure the confidentiality of the data. Taking one step further, for the first time in the mental health domain, the CLPSych 2021 shared task has introduced a more secure approach where sensitive data is stored within a secured

network so that the researchers can work with the data in a secured environment (MacAvaney et al., 2021).

We obtained the ethics approval certificate for CLPSych 2015 and the UMD datasets and provided a signed data usage agreement to acquire the SMHD dataset. The WASSA-2017 dataset is publicly-available for research purposes. During our research, we have given thorough considerations to these ethical facets and have adopted strict guidelines to ensure the anonymity and privacy of the data. Following the ethical guidelines and best practices proposed by Benton, Coppersmith, and Dredze (2017) and Resnik et al. (2021), we implemented strict hardware and software security measures and best practices such as: allowing only the researchers listed in the certificate of ethics approval to access data, storing data in a secured server with strict accessibility guidelines and not including examples of the text published by at-risk users in any publication. Also, our research does not involve any intervention and has focused mainly on the applicability of machine learning models in determining users susceptible to mental disorders and suicide ideation.

1.6 Contributions

- We used limited social media data to train a multi-task learning model with hard parameter sharing and automatically-calculated auxiliary inputs (e.g., emotion categories, age, gender) to detect mental health conditions of users (i.e., if the user is neurotypical or shows signs of PTSD or depression).

- To the best of our knowledge, in the mental illness and suicide ideation detection domain, we are the first to use multi-task learning with both soft and hard parameter sharing to explore the impact mental disorders detection have on

suicide ideation detection, and also the impact suicide ideation detection has on mental illness detection using two datasets with different distributions (i.e., without combining the datasets into a single dataset).

- To the best of our knowledge, we are the first to explore the impact and the level of impact single and multiple mental disorder detection have on suicide ideation detection, and also the impact and the level of impact suicide ideation detection has on the particular mental illness detection.

- To the best of our knowledge, we are the first to use data from two different social media platforms (i.e., Twitter and Reddit social media platforms) with different distributions to identify if knowledge can be successfully transferred between suicide ideation and mental illness detection tasks. The experiments were conducted using different permutations of the given data containing different percentages of users diagnosed with PTSD or depression. In the process, we identified the impact different mental disorders have on suicide ideation detection.

1.7 Thesis Structure

The rest of the thesis is organized as follows:

- Chapter 2 provides an overview of the related theoretical background of natural language processing and machine learning methods used in the research. The chapter also lists the reasons why specific state-of-art methods are not used in our research. The last section of the chapter lists related work in mental illness and suicide ideation detection that uses natural language processing and machine learning (including deep learning) methods.

- Chapter 3 explains the datasets we have used in our research. The chapter lists the details about the four datasets used to predict emotion, mental disorders and suicide ideation. We have included in-depth information about the datasets and how we prepared the data for our research so that we could focus mainly on our proposed solution and its outcomes in the relevant chapters (i.e., chapters 4, 5 and 6). The chapter also includes the details about data pre-processing and exploratory analysis used to identify the auxiliary features for training our models.

- Chapter 4 explains the proposed solution that uses multi-task learning with hard parameter sharing to predict neurotypicals and users with mental disorders (i.e., depression and PTSD). The chapter also explains the generation of the auxiliary inputs (e.g., emotion categories) and provides a detailed explanation of the conducted experiments and their results. Finally, we discuss the results and compare them with state-of-the-art solutions from related research.

- Chapter 5 details the proposed solution that uses multi-task learning with hard and soft parameter sharing to identify individuals with suicide ideation and mental disorders. The chapter lists the details about the model architecture, experiments and their results, followed by a discussion and a comparison with state-of-the-art solutions from related research.

- Chapter 6 discusses using the proposed architecture in chapter 5 to explore knowledge transfer between tasks using data from two different social media platforms (i.e., Twitter and Reddit) and distributions. The chapter lists details about the experiments and results, followed by a discussion on the obtained

results. Finally, the results are compared with state-of-the-art solutions from related research.

- Chapter 7 concludes our research by summarizing our findings. Also, it includes the direction of our future research in the domain of mental illness and suicide ideation detection.

1.8 Publications

Listed below are the publications related to the area of research:

- Inkpen, D., Skaik, R., Buddhitha, P., Angelov, D., & Fredenburgh, M. T. (2021). uOttawa at eRisk 2021: Automatic Filling of the Beck's Depression Inventory Questionnaire using Deep Learning. In Conference and Labs of the Evaluation Forum.

- Kirinde Gamaarachchige, P., & Inkpen, D. (2019). Multi-Task, Multi-Channel, Multi-Input Learning for Mental Illness Detection using Social Media Text. In Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019) (pp. 54–64). Hong Kong: Association for Computational Linguistics.

- Husseini Orabi, A., Buddhitha, P., Husseini Orabi, M., & Inkpen, D. (2018). Deep Learning for Depression Detection of Twitter Users. In Fifth Workshop on Computational Linguistics and Clinical Psychology (pp. 88–97). New Orleans: Association for Computational Linguistics.

- Jamil, Z., Inkpen, D., Buddhitha, P., & White, K. (2017). Monitoring Tweets for Depression to Detect At-risk Users. In Proceedings of the Fourth Workshop

on Computational Linguistics and Clinical Psychology (pp. 32–40). Vancouver: Association for Computational Linguistics.

- Buddhitha, P., & Inkpen, D. (2017). Topic-Based Sentiment Analysis. In J. A. Lossio-Ventura & H. Alatrista-Salas (Eds.), Information Management and Big Data (pp. 95–107). Cham: Springer International Publishing.

- Buddhitha, P., & Inkpen, D. (2015). Dependency-based Topic-Oriented Sentiment Analysis in Microposts. In J. A. Lossio-Ventura & H. Alatrista-Salas (Eds.), Proceedings of the 2nd Annual International Symposium on Information Management and Big Data - SIMBig 2015 (Vol. 1478, pp. 25–34). Cusco: CEUR-WS.org.

Chapter 2

Background and Related Work

2.1 Background

2.1.1 Kernel-based Methods and Feature Engineering

Kernel Methods

Some of our early experiments were based on kernel methods and explicitly used the Support Vector Machine (SVM) classification algorithm. In those experiments, we mapped a given set of examples to known labels. As illustrated in figure 2.1 (Cortes & Vapnik, 1995), SVM transforms the data into a high dimensional space where the data points belonging to different groups are separated using a surface known as a hyperplane. To optimize the class separation, the decision boundary is calculated by maximizing the distance between each point belonging to different classes and the hyperplane (Bishop, 2006).

To overcome the computational cost of distance calculation between each data point and the hyperplane, a more practical approach was introduced using a method called kernel function that calculates the distance between data point pairs. The kernel function compares the distances between two points in the initial and the target

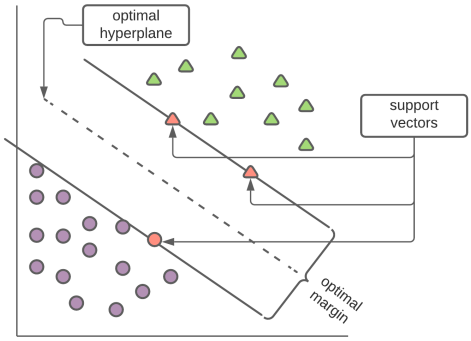


Figure 2.1: Support vector classifier with soft margins in a two dimensional space

space (Burges, 1998).

We used the SVM algorithm to calculate a baseline to be compared against the proposed approach in detecting mental disorders using multi-task learning. We tuned the parameters: *gamma* and C to obtain the optimum classification results. The *gamma* value is used to measure the width of the Gaussian kernel, where increasing the value will inherently increase the model complexity. The parameter C is used as a regularization parameter to control the importance of each data point, where increasing the given value for C demonstrates a positive correlation to the level of influence each data point has on the trained model.

Term Frequency-Inverse Document Frequency

With the Support Vector Machine classification algorithm, we used Term Frequency-Inverse Document Frequency (TF-IDF) scores as manually engineered features. TF-IDF is the product of term frequency and inverse document frequency which can be used in downweighing the frequently used words. We obtain the TF-IDF score as below:

$$tf\text{-}idf_{(t,d)} = tf_{(t,d)} \times log_{10}(\frac{N}{1 + df_{(d,t)}}) \tag{2.1}$$

$tf_{(t,d)}$ = number of times the term t occurs in document d

N = total number of documents

$df_{(d,t)}$ = total number of documents that contain the term t

2.1.2 Deep Learning for Natural Language Processing

For years, Natural Language Processing research has focused more on classical machine learning methods where feature transformation exists within one or two layers. However, in recent years, researchers have focused more on Deep Neural Network architectures using dense vector representations. Apart from the availability of large datasets, one of the key reasons for deep learning methods to be more effective is the representation of words as dense vectors. To obtain a dense representation of words when trying to detect mental disorders, we used word embeddings trained on word2vec (Mikolov et al., 2013) and fastText (Joulin et al., 2017) algorithms. Word embeddings are floating-point dense vectors (i.e., low dimensional) that are learned using text data. Unlike words represented as sparse one-hot vectors, word embeddings contain more information in a low-dimensional vector space. After several preliminary experiments, we selected fastText for our research as it considers words at their morphological level. In addition to the experimental outcome, we decided to use fastText because we could obtain a more meaningful representation by expressing a word as a vector constructed out of the sum of several vectors for different parts of the word (Bojanowski et al., 2017). The word representation is constructed as follows:

$$S(w, c) = \sum_{g \in G_w} Z_g^T V_c \tag{2.2}$$

w = given word

G_w = set of n-grams in the word w

g = individual n-grams

Z_g = vector representation of the n-gram

V_c = context vector

In order to obtain a vector representation for the word w, first, it will be formed into a bag of character n-grams. For each character n-gram, the dot product is calculated between the character n-gram vector and the context vector, and the sum of all the dot products will represent the word w given the context c. One of the key reasons that make the fastText algorithm better than the word2vec algorithm is the capability of obtaining vector representations for out-of-vocabulary words.

Even though we conducted experiments using pre-trained embeddings, the best models obtained for both "multiple mental disorders" and "suicide ideation and mental disorder" detection using multi-task learning were by instantiating the embedding layer weights with random numbers. One of the reasons for the low measurements when detecting mental illnesses using Twitter data could be due to the reason that we used fewer data to train our embeddings, and also, we extracted the most similar words for out of vocabulary words when generating the embedding matrix to be used in the embedding layer. These words might not represent the intended meaning within the given context. Concerning the suicide ideation detection tasks, we did not train custom embeddings on suicide data. Instead, we used pre-trained embeddings

made available in the flairNLP (Akbik et al., 2018) natural language processing library. The proposed suicide ideation and mental illness detection models were trained using different pre-trained word embeddings: fastText (Grave et al., 2018), GloVe (Pennington et al., 2014), Byte-Pair Embeddings (Heinzerling & Strube, 2018), Character Embeddings (Lample et al., 2016) and stacked embeddings where a few of the previously mentioned embeddings were combined to form a single embedding layer (Akbik et al., 2018). Our research did not put emphasis on different word embeddings since our objective is to identify if mental disorders positively impact suicide ideation detection given the social media data. Once we identified the positive impact mental disorders have on detecting suicide ideation using randomly initialized embeddings and produced comparable outcomes against state-of-the-art in the domain, we did not further investigate creating embeddings using different architectures but rather focused on fine-tuning the randomly initialized embeddings. With experiments on different embedding dimensions, the optimum performance when detecting multiple mental disorders was achieved using a dimension of 100, and a dimension of 300 produced the best results when detecting suicide ideation and mental disorder. We will further investigate using more informative embeddings, including transformer-based architectures, to obtain effective mental illness and suicide ideation predictions in our future work.

Transformer based Architectures

Despite the recent groundbreaking results in classification tasks obtained using transformer-based (i.e., transformers proposed by Vaswani et al. (2017)) architectures, such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et

al., 2019), RoBERTa (A Robustly Optimized BERT Pretraining Approach) (Liu et al., 2019), we did not use such methods due to several implementation limitations. One of the limitations BERT has is its inability to process long sequences of text. BERT only considers up to 512 tokens, that is, including the two special tokens [CLS] (classification embedding) and [SEP] (sentence separator). Most of the research with sequence lengths above the given limit tends to chunk the sequences and combine the predictions coming from a transformer-based model (Pappagari et al., 2019). Chunking long sequences introduce drawbacks when calculating the attention, which will be localized into the chunked sequence of tokens. Several architectures have been proposed to overcome such limitations. Beltagy et al. (2020) introduced Longformer, which reduces the negative impact self-attention has on long sequences by introducing a combination of local and global attention. The authors managed to process sequence lengths up to 4,096 tokens.

Kitaev et al. (2020) efficiently processed documents with sequences of up to 64,000 tokens by calculating attention using locality-sensitive hashing instead of the dot-product and efficiently storing the activations using reversible residual layers. Another limitation was the excessive need for computational resources to implement transformer-based architectures, specifically when using longer sequences. Even with effective and efficient resource management methods introduced by Reformer (Kitaev et al., 2020), we still found it resource-intensive (i.e., given the limited Graphics Processing Units) given the research data for mental illness and suicide ideation detection. We conducted several preliminary experiments by chunking long sequences into manageable blocks of tokens and processing them through several transformer-based architectures such as BERT, RoBERTa and DistilBERT (Sanh et al., 2020). The

output from the transformer-based architecture was stacked horizontally to construct the complete sequence, and the resulting matrix was fed into a Convolution Neural Network. Due to the resource intensiveness, we had to progressively freeze several layers so that the model could train without running out of memory. The model's predicted outcome was substantially poor, so we did not report the results in the thesis. Nevertheless, we will be conducting additional research on the architecture to discover further enhancements that could be applied to improve mental illness and suicide ideation detection in future work.

Convolutional Neural Networks(CNN)

Even though the CNN architecture (LeCun et al., 1989) is mainly used to extract features from data that can be identified with a spatial relationship, we used Convolution Neural Networks in our classification tasks to extract ordered relationships. The Convolutional Neural Networks use the mathematical convolution operation instead of the general matrix multiplication approach used in many neural network architectures. The CNN models output a feature map by applying the convolution operation on the input using a matrix known as the kernel. By convention, the size of the kernel used in the convolution operation is smaller than the dimensions of the input tensors, which allows identifying features of smaller dimensions. Compared to many other neural network architectures such as Recurrent Neural Networks (RNN) and Multilayer Perceptrons, the CNNs have the advantage of learning fewer parameters, which will considerably reduce the computational overhead. Also, the architecture allows sharing parameters among different parts of the given text by reusing the feature detector. For CNNs to work with text, the input can be considered a three-dimensional object

where the shape consists of the number of samples, input sequence length and the number of features (i.e., the embedding dimension). The sequence length represents a temporal axis, where based on the convolution window size, certain temporal features can be extracted. Unlike when CNNs are used to identify features within the images where a convolution window of size $n \times n$ extracts n^2 features, a window of size n will only be able to extract n number of features from a one-dimensional textual object. It is important to highlight that even though a certain level of temporal features can be extracted from the CNNs, theoretically, it does not have the equivalent impact of Recurrent Neural Network based architectures. In addition to the convolution operation, the CNN network layers are regularly associated with the pooling layer, which is used to reduce the representation of the model after being transformed using a nonlinear activation function. The functionality could make the trained model more robust by making the representation consistent over minor changes to the input (Goodfellow et al., 2016). Our experiments identified that a more stable model could be trained using a global max pooling layer instead of more widely used max pooling or average pooling layers. Figure 2.2 represents a multi-channel (i.e., using different kernel sizes) convolution neural network architecture (Y. Kim, 2014) with a global max pooling layer. The CNN model uses multiple filters where each different group of filters has three different kernel sizes (i.e., 1, 2 and 3), which in our case resemble different n-gram sizes. The activated output after the convolution operation is sent through a global max pooling layer to downsample the transformed output. The outputs from the pooling layers will then be concatenated and either transformed further using different types of neural network layers, or submitted to a softmax function if the end task is to do multi-class classification, or to a sigmoid function for binary classification.

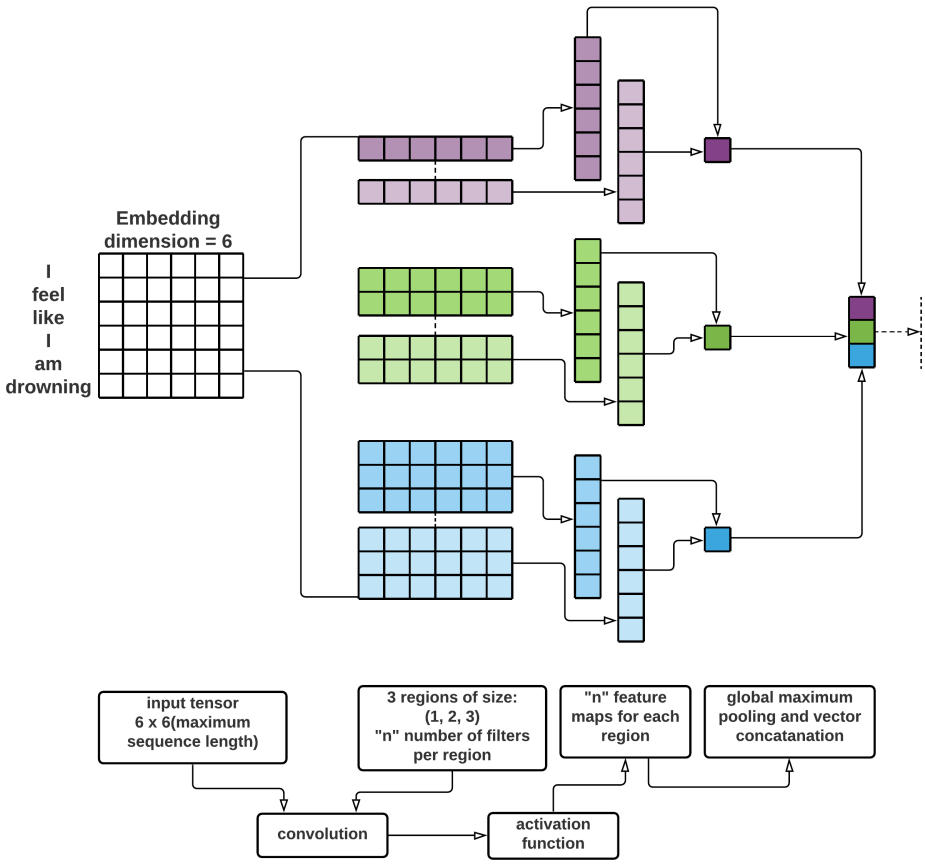


Figure 2.2: Multi-channel CNN architecture with global max pooling.

Multi-Task Learning

Multi-Task Learning (MTL) learning aims to generalize the primary task by sharing representations of related tasks (Caruana, 1997). MTL approaches can be categorized mainly into two types, based on how the parameters are being shared. One approach is hard parameter sharing, where model weights are shared between the tasks (Caruana, 1997). According to figure 2.3, the end tasks can share one or more hidden layers, while features unique to the specific task will be learned using the task-specific layers.

The second approach is soft parameter sharing, where we conducted several

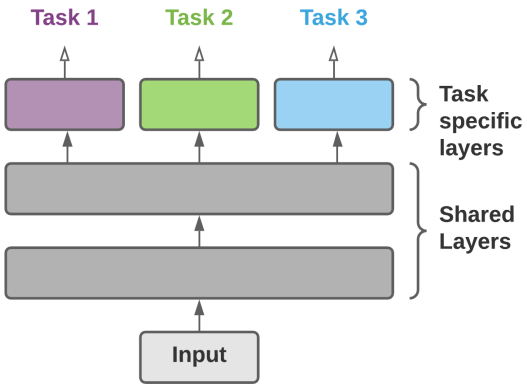


Figure 2.3: Multi-task learning with hard parameter sharing.

experiments to identify users susceptible to mental disorders (i.e., MTL task one) and suicide ideation (i.e., MTL task two). When using this approach, each task is trained using its subnetwork (as shown in figure 2.4) without sharing any parameters between the layers. Even though the parameters are not shared, they are regularized between the layers of the sub-models to obtain similarity (Ruder, 2017). The parameter regularization is enforced by using a custom loss function.

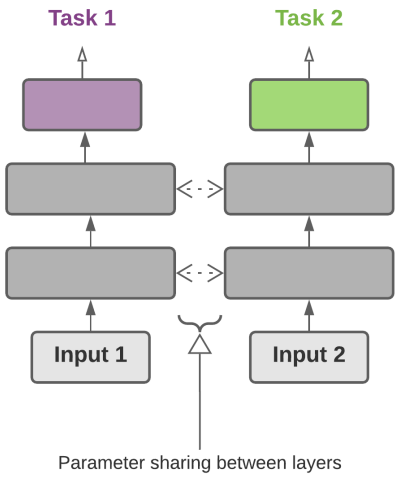


Figure 2.4: Multi-task learning with soft parameter sharing.

When conducting multi-task learning using suicide ideation and mental disorder detection data, we identified that the best-performing model is a combination of soft and hard parameter sharing, also referred to as mixed parameter sharing (Raaijmakers, 2021). As shown in figure 2.5, the mixed parameter sharing model combines both the before mentioned approaches (i.e., according to figure 2.3 and figure 2.4). Similar to soft parameter sharing, a certain level of constraints (i.e., by using a custom loss function) were enforced on layers within the subnetworks to minimize the distance of the weight vectors. In addition to soft parameter sharing, both end tasks will share certain layer representations.

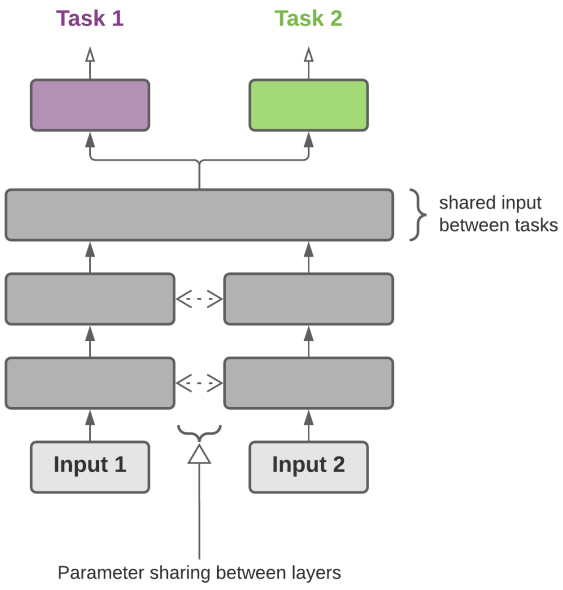


Figure 2.5: Multi-task learning with hard and soft parameter sharing.

As our research findings are on multiple mental disorder detection (chapter 4) and suicide ideation and mental disorder detection (chapter 5), we conducted research using all the above-mentioned multi-task learning architectures. Hard parameter sharing being the most widely used approach (Ruder, 2017), we used it for our experiments in

detecting multiple mental disorders. With comparable results to previous research, we did not further investigate the use of soft parameter or mixed parameter sharing to detect multiple mental disorders. However, when conducting research on both suicide ideation and mental disorder detection as multiple tasks in an MTL environment, we discovered that the mixed parameter sharing approach outperforms both the soft and hard parameter sharing architectures.

The Objective Function

In multi-task learning, one of the main concerns is how to construct the objective function or the loss function to minimize the loss while improving the convergence of each task in the model. We used binary cross-entropy as the loss function when predicting multiple mental disorders (i.e., PTSD and Depression). Because the loss function must be minimized across several tasks, binary cross-entropy calculated for each task was accumulated into the final loss.

$$BCL = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1 - y_i)\, log(1 - \hat{y}_i)\right] \tag{2.3}$$

i = number of samples from 1 to N

y = the desired output

$\hat{y}$ = individual prediction

Binary cross-entropy (see the formula 2.3) loss is calculated by taking into consideration the assigned class and the predicted probabilities. For example, if the desired output is zero, $y \log \hat{y}$ will be evaluated to zero, and in order to minimize the loss, $\hat{y}$ must be closer to zero. Similarly, if the desired output is one, $(1 - y)\, log(1 - \hat{y})$ will

be zero and, in order to minimize the loss, $\hat{y}$ must be closer to one. As illustrated in formula 2.3, the summed loss (i.e., for all the inputs) is averaged by dividing from the number of inputs (i.e., N). When training a neural network model, a mini-batch is randomly selected rather than using a single sample at a time. The loss is calculated on this mini-batch, and in order to minimize the loss given a differentiable function, gradients will be calculated on the identified loss (i.e., concerning the trainable network parameters). The process of calculating the gradients is known as backpropagation or the backward pass. A pass through the entire training dataset, which is a collection of many mini-batches, is known as an epoch, and the losses calculated on the mini-batches are averaged at the end of each epoch.

When researching mental illness and suicide ideation detection, we used a custom loss function that summed the output from "categorical cross-entropy" loss (CL) and "mean squared error" loss. As shown in formula 2.4, categorical cross-entropy takes the desired output vector (in our case, one-hot encoded true labels) and multiplies it with the logarithmic values of the predicted softmax output $\hat{y}$. As illustrated in formula 2.4, the summed loss (i.e., for all the inputs) is averaged by dividing from the number of inputs (i.e., m). Cross entropy is calculated for each class (i.e., from $j = 1$ to c), and in our case, it will be two classes representing the positive and the negative labels for each task. We did not use binary cross-entropy to accommodate the calculation of the model averaging ensemble approach.

$$CL(\hat{y}, y) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{c}[y_j^{(i)}\ \log(\hat{y_j}^{(i)})] \tag{2.4}$$

i = number of instances from 1 to m

j = number of classes from 1 to c

y = the desired output

$\hat{y}$ = individual prediction

Mean squared error (MSE) loss (formula 2.5) is generally used to calculate the distance between the target and predicted vectors so that through optimization, the distance or the loss can be reduced. We used mean squared error to calculate the difference between the intermediate weight vectors generated from each subnetwork, where the subnetworks are the segments processing the tasks of suicide ideation and mental illness detection. Used within the custom loss function, MSE regularizes the parameters in the subnetwork layers.

$$MSE = -\frac{1}{N}\sum_{i=1}^{N}(V_{T1}^{(i)} - V_{T2}^{(i)})^2 \tag{2.5}$$

i = observation number from 1 to N

N = total number of observations

V_{T1} = vector one (in our case, intermediate weight vectors from either the suicide ideation or mental illness detection subnetwork)

V_{T2} = vector two (in our case, intermediate weight vectors from either the suicide ideation or mental illness detection subnetwork)

Tuning Deep Neural Networks

The most critical and challenging part when training a deep neural network is selecting hyperparameters. Hyperparameters are parameters that are manually configured,

unlike parameters that get trained through backpropagation. Due to resource limitations, we manually searched for the best hyperparameters (around the most promising values) and did not use exhaustive search strategies, such as Bayesian optimization or Hyperband (Li et al., 2018). The sections below highlight the hyperparameters that had a significant impact on enhancing the model performances of the proposed solutions.

Batch Size Before training a deep neural network model, it is essential to initialize the hyperparameters with values that have been proven to work well based on the model being developed and the data being used. For example, selecting a smaller batch size to start the training could be more effective rather than selecting a bigger batch size. Similar to the empirical study conducted by Masters and Luschi (2018), the best results for the experiments were obtained for batch sizes equivalent to or less than 32. Throughout our experiments, we identified that smaller batch sizes produced better results compared to using a large bath size. Smaller batch sizes will inevitably have a less memory footprint, and our research have also shown that the optimal performances could be obtained when using a mini-batch of size 8 (i.e., when using MTL for "mental illness detection" and "suicide ideation and mental disorder detection") and increasing it could harm the models' performance.

Learning Rate: The learning rate could be considered one of the key hyperparameters to be tuned as it decides on the phase in which the algorithm should learn. In other words, the calculated gradient values will be multiplied by the learning rate during the backward pass. The network weights will be adjusted according to the calculated value. If the network weights were adjusted accordingly, the newly calculated loss

would get reduced within the next forward pass. A larger learning rate could surpass the minimum loss while having a lower learning rate could take longer to arrive at the minimum loss. It is crucial to identify the correct learning rate so that the objective function will arrive at a global minimum. We discovered that having a learning rate of 0.001 produced better results for all the experiments we conducted than using a smaller or a larger learning rate.

Activation Functions: A neuron in a neural network performs a linear operation (i.e., before applying linear or nonlinear activation) which consists of a dot product (i.e., between the input and the weight parameter) and an addition (i.e., the bias). Similarly, a collection of neurons, known as a layer, will also perform a linear transformation. In certain situations, it might not be enough to perform only a linear transformation because nonlinear transformations could define a better decision boundary. For that reason, it is vital to perform nonlinear transformations on the input data and only when necessary to use the linear transformation (e.g., sigmoid function in the output layer). During our research, we experimented with different activation functions in the hidden layers and identified the ones that produced better results given the task. Listed below are a few of the activation functions that we have tested.

The sigmoid function is used in the output layer for binary classification, while the softmax activation is used when conducting multiclass classification. The sigmoid function (figure 2.6 (a)) will produce a value that can be interpreted as a probability within the range of 0 and 1. When detecting "multiple mental disorders" and "suicide ideation and mental disorder", we achieved the best results using the ReLU activation function. As illustrated in figure 2.6 (b), the ReLU (Rectified Linear Unit) function will output zeros when given a negative input (Nair & Hinton, 2010). To overcome the

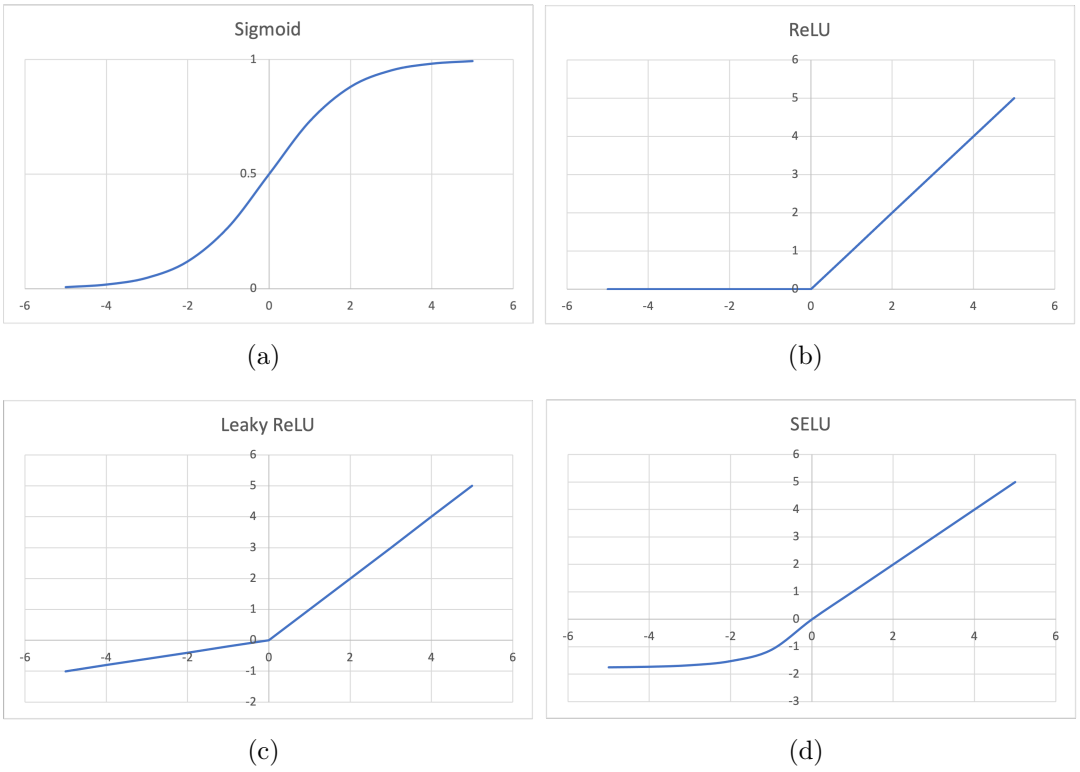


Figure 2.6: (a) Sigmoid (b) ReLU (c) Leaky ReLU (d) SELU, activation functions.

issue of neurons producing zeros, Maas et al. (2013) introduced Leaky ReLU (figure 2.6 (c)). The Leaky ReLU function uses the hyperparameter α to indicate how much of a leak or the slope of the function should be when the input (i.e., z according to formula 2.6) is negative. The best results for our experiments when detecting suicide ideation and mental disorder (i.e., when using Twitter data as the input for detecting mental illness) were achieved when we used the Leaky ReLU activation with a hyperparameter α value of 0.2. It was identified that using a higher value for α tends to produce better results compared to a more commonly used value mentioned by Maas et al. (2013), which is 0.01 (Xu et al., 2015). Several experiments were also conducted by using the SELU (Scaled Exponential Linear Unit) activation (figure

2.6 (d)), but it could not surpass the results obtained using ReLU and Leaky ReLU in our case. Klambauer et al. (2017) have shown that using the SELU activation function within a stack of hidden dense layers will normalize the network data. For the SELU activation to be effective, certain conditions must be applied (e.g., using LeCun Normal to initialize the layer weights).

$$LeakyReLU(z) = \max(\alpha z, z) \tag{2.6}$$

Model Overfitting and Underfitting: Model overfitting and underfitting are vital issues in machine learning that needs significant attention. The main reason for dedicating a separate subsection to discuss model overfitting and underfitting and a brief explanation of the steps we have taken to reduce their impact, especially for overfitting, is its considerable impact on generalization, primarily when used for suicide ideation and mental illness detection tasks. Model overfitting is when there is a low bias and high variance, and underfitting is when there is a high bias and high variance. Overfitting implies that the model has not generalized and yet has parameterized well onto the training data. There could be many reasons that lead to overfitting, such as having a complex model and limited data, training and validation data coming from different distributions and class imbalance between the training and validation datasets, to name a few. The following regularization methods can be implemented to overcome the negative impact of overfitting.

Early stopping: Once the model has stopped learning (e.g., the validation loss has not improved over a given number of epochs), the learning can be stopped, and the model weights that produced the least error on the validation data can be returned.

Dropout: Probabilistically dropping neurons from the network during training so

that the model will not memorize all the features from the training data (Srivastava et al., 2014).

Reduce Network Complexity: Reduce the network complexity by removing layers or having fewer hidden units. It is important to reduce the network complexity, especially when having a smaller number of data points, where otherwise there will be a sufficient number of parameters within the network to memorize all the features extracted from the limited training data (Goodfellow et al., 2016). If adequate training data is available, it is preferable to add more layers to the network so that the model can effectively learn the necessary features.

Weight Regularization: Weight regularization is used to penalize the model during the network optimization phase. Weight regularization makes the network weights smaller, and as a result, the distribution of the weights will be in a specific range. Two of the main regularization techniques applied when implementing weight regularization are L1 and L2 regularization. Both methods will add a cost to the network loss for having larger weights (Chollet, 2017). Weight regularization can be applied on each layer and different parameters such as layer weights or the kernel, layer bias, and layer output (i.e., after applying the activation function) (Chollet, 2017). Given limited data, we achieved better model generalization by applying L1 and L2 regularizers on the kernel and L2 regularization on the activated layer outputs.

2.1.3 Ensemble Methods

Training different candidate networks could produce different outcomes, yet selecting the best-performing model to predict on unseen data might not generalize well (Bishop, 1995). Even multiple runs of the same neural network model do not guarantee to

reproduce the same results as expected. The variance in the predicted results can be due to the random initializations of the weights. In order to reduce the randomness between predictions, one can make the random initialization of the weights and biases more static during different runs. Even static initializations do not guarantee the same output, especially when training the model in a single or multi-GPU environment. To further reduce such randomness and in addition to setting a random seed, an ensemble learning approach can be used (Brownlee, 2018). An ensemble approach can combine different predictions and generate the final output based on an ensemble strategy. Several ensemble strategies were proposed in Brownlee (2018), such as: "multiple training run ensemble", "snapshot ensemble", "model averaging ensemble", "weighted average ensemble" to name a few. We used the model averaging ensemble for our experiments, where prediction probabilities from individual models are combined before taking the argmax to generate the predicted class label. We used this approach only with the suicide ideation and mental illness prediction experiments and not when using multi-task learning to predict multiple mental disorders.

2.1.4 Evaluation Metrics

Selecting the correct evaluation metrics is vital when training any machine learning model. In our research, we used the evaluation metrics recommended by the researchers from whom we obtained the data. We used the following metrics to evaluate the model performances.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{2.7}$$

$$Recall(Sensitivity) = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{2.8}$$

$$F1 = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \tag{2.9}$$

The area under the receiver operating characteristic curve (ROC AUC) can also be considered as a better metric to evaluate model performances, especially when it is required to identify the overall model performances for different classification thresholds that separate the true positive rate from the false positive rate. The ROC graph plots the True Positive Rate (Sensitivity) against the False Positive Rate (1 – Specificity), and the AUC, as the name states, measures the area under the ROC curve. The true positive rate is also known as Recall. False positive rate calculates the ratio of incorrect classification of negative samples as positive. A ROC AUC value greater than 0.5 suggests that the trained model performs better than when the true positive rate equals the false positive rate.

2.2 Related Work

2.2.1 Social Media and Self-Disclosure

As social media has become an integral part of one's day-to-day life, it will be insightful to identify to what extent an individual has disclosed her/his personal information and whether accurate and sufficient information is being published to determine whether or not a person has a mental disorder. Self-disclosure means to reveal the unknown facts about one's self so that they become shared knowledge, or, in other words, the "process of making the self known to others" (Joinson & Paine, 2007). Considering the Twitter platform, rather than sharing feelings about depression, users are more likely to self-disclose detailed information about their treatment history (Park et al.,

2013). Compared with the CES-D (Center for Epidemiologic Studies Depression Scale) score, a strong correlation was identified between the particular score obtained by a user and specific demographics and sentiment predictors (e.g., anger, anxiety, sadness, and causation) (Park et al., 2013). The same level of self-disclosure can be identified in the Reddit forums (Balani & De Choudhury, 2015) and specifically by users with anonymous accounts (Pavalanathan & De Choudhury, 2015). Also, it was identified that personality traits and meta-features such as age and gender could have a positive impact on the model performances when detecting users susceptible to PTSD and depression (Preot et al., 2015). Similarly, we have also identified that age and gender as auxiliary inputs when detecting multiple mental disorders in a multi-task learning environment have enhanced the model predictability.

2.2.2 Mental Illness Detection

Text extracted from social media platforms such as Twitter, Facebook, Reddit and other similar forums has been successfully used in various natural language processing (NLP) tasks to identify users with different mental disorders and suicide ideation. Social media text was used to classify users with insomnia and distress (Jamison-Powell et al., 2012; Lehrman et al., 2012), postpartum depression (De Choudhury et al., 2013b, 2013a, 2014), depression (Resnik, Armstrong, Claudino, Nguyen, Nguyen, & Boyd-graber, 2015; Resnik et al., 2013; Schwartz et al., 2014; Tsugawa et al., 2015), Post-Traumatic Stress Disorder (Coppersmith et al., 2014b, 2014a), schizophrenia (Loveys et al., 2017) and many other mental illnesses such as Attention Deficit Hyperactivity Disorder (ADHD), Generalized Anxiety Disorder, Bipolar Disorder, Eating Disorders and obsessive-compulsive disorder (OCD) (Coppersmith, Dredze,

Harman, & Hollingshead, 2015). An overview of the literature in detecting mental disorders using social media data is mentioned in Table 2.1. The table identifies the individual research, its objective, the different features, and the methods used to generate those features and the main algorithm or approach used to determine individuals susceptible to one or more mental disorders.

Mentioned below are the abbreviations used in Table 2.1: LIWC (Linguistic Inquiry and Word Count), TF-IDF (Term Frequency - Inverse Document Frequency), LDA (Latent Dirichlet Allocation), SVM (Support Vector Machine)

Literature	**Objective**	**Features**	**Methods**
Jamison-Powell et al. (2012)	Insomnia (Twitter data)	LIWC features	Thematic and content analysis
Lehrman et al. (2012)	Distress (online forums)	Sentence length, word polarities, affective states, part of speech	Naïve Bayes, decision trees
De Choudhury et al. (2013b)	Postpartum depression (Twitter data)	Linguistic style, emotional expressions and posting patterns	SVM
Tsugawa et al. (2015)	Depression (Twitter data)	User behavioral characteristics, egocentric social graph, depression language, emotion, linguistic style, bag-of-words and word frequencies, topic modeling	SVM
De Choudhury et al. (2013c)	Depression (Twitter data)	Features from the Twitter posts, users, and ego-network, post time	SVM
Schwartz et al. (2014)	Depression (Facebook data)	n-grams, LIWC features, sentiment, LDA topics	Linear regression
Coppersmith et al. (2014a)	PTSD (Twitter data)	Character n-grams, LIWC features	SVM

Literature	Objective	Features	Methods
Coppersmith et al. (2014b)	PTSD, depression, bipolar disorder, seasonal affective disorder (SAD) (Twitter data)	Language models, LIWC and user behavioral data (e.g., tweet rate)	Log-linear classifier
Coppersmith, Dredze, Harman, and Hollingshead (2015)	Ten mental illnesses: Attention Deficit Hyperactivity Disorder (ADHD), Generalized Anxiety Disorder, Bipolar Disorder (Twitter data)	LIWC features, open-vocabulary, character n-grams	Classifier not mentioned
Mitchell et al. (2015)	Schizophrenia (Twitter data)	LIWC features, lexicons, LDA topics, brown clustering, perplexity values, character n-grams	SVM
Resnik, Armstrong, Claudino, and Nguyen (2015)	Depression, PTSD (Twitter data)	Supervised topic modeling, supervised anchor algorithm, TF-IDF	SVM
Preotiuc-Pietro et al. (2015)	Depression, PTSD (Twitter data)	Bag-of-words, topics derived from clustering methods, meta data	Ensemble of classifiers (logistic regression and SVM)
S. M. Kim et al. (2016)	Post severity in mental health related forums	TF-IDF, post embeddings	Ensemble of classifiers (logistic regression and SVM)

Literature	Objective	Features	Methods
Malmasi et al. (2016)	Post severity in mental health related forums	Lexical (e.g. character n-grams, word n-grams), syntactic features (e.g., part-of-speech n-grams), meta data	Random Forest meta-classifier

Table 2.1: Mental illness detection related literature summary

With the advancements in neural network-based algorithms, more research has been conducted successfully in detecting mental disorders, despite the limited amount of data. Kshirsagar et al. (2017) have used recurrent neural networks with attention to detect social media posts resembling crisis. Husseini Orabi et al. (2018) demonstrated that using convolution neural network-based architectures produces better results compared to recurrent neural network-based architectures when detecting users susceptible to depression.

Intuitively, sharing representations between different but related tasks (in our case, detecting PTSD and depression) is well adapted for mental illness detection where certain mental disorders share specific characteristics (Benton, Mitchell, & Hovy, 2017; Coppersmith, Dredze, Harman, & Hollingshead, 2015). Benton, Mitchell, and Hovy (2017) used multi-task learning to detect users susceptible to suicidal risk and mental illnesses and highlighted that using multi-task learning to detect suicide risk and mental disorders produces significant results over single-task learning methods. Even though our work could not be directly compared with Benton, Mitchell, and Hovy (2017) due to the different datasets being used, we can identify that our model has produced competitive results, especially when comparing the AUC score for detecting users with PTSD and depression. At the end of each chapter 4, 5 and 6, we have

compared our results against the state-of-the-art research conducted in the same or related area.

2.2.3 Suicide Ideation Detection

Research shows that in addition to the impact mental disorders have on suicide ideation or suicide thoughts, certain mental disorders such as PTSD, bipolar and substance use disorders are strong indicators of suicide attempts (Nock et al., 2009). Considering these factors, we can highlight the importance of predicting mental disorders and suicide ideation which could inherently reduce the number of suicide attempts. Due to certain limitations in the clinical approaches (as discussed in chapter 1), researchers have identified the importance of early detection of suicide ideation using machine learning methodologies. In recent years, more research has been conducted using social media data due to effective data collection methods based on matching self-reported diagnosis content with predefined phrases (Coppersmith et al., 2014b).

With respect to feature engineering methods, Coppersmith et al. (2016) have demonstrated a quantifiable approach to determine whether an individual shows signs of suicide risk in the present time and also whether there will be a risk of suicide in the future. In analyzing the emotional state of a user before and after the suicide attempt and even in comparison to the control group, the authors have identified: the use of a lower number of emoticons and emojis, talking about suicide after the attempt rather than before, the use of self-focused language, an increase in sadness before the suicide attempt, an increase in anger and sadness after the attempt, a decrease in the level of fear and disgust after the attempt, a decreased number of tweets resembling loneliness, and a decline in the overall number of tweets but an increase in

the number of emotional tweets. Even with limitations such as the underrepresentation of the sample population, this exploratory analysis demonstrated the possibility of applying technology for screening users to determine the risk of a suicide attempt. Similar to mental illness detection, features based on lexical, sentiment and emotional characteristics (Burnap et al., 2015), communication network (Colombo et al., 2016) topic modelling (Abboute et al., 2014; Huang et al., 2015), interpersonal awareness and interaction (De Choudhury et al., 2016) were identified as contributors in identifying individuals with suicide ideation.

In recent years more researchers have opted to use deep learning methods to predict users having suicide ideation, especially with the availability of more data through shared tasks such as CLPSych 2019 (Zirikly et al., 2019) and CLPSych 2021 (MacAvaney et al., 2021). The CLPSych 2019 shared task focused on predicting the level of suicide risk demonstrated by users who have published posts on the Reddit social media platform. The shared task included three subtasks where each subtask used different data combinations from the Reddit social media platform. For "Task A", data from the subreddit SuicideWatch are used to predict the users' level of suicide risk. In addition to the "Task A" data, "Task B" has used data published elsewhere in Reddit by the users filtered from the SuicideWatch subreddit. For "Task C", the participants were given Reddit data other than from the SuicideWatch subreddit. For each of the before mentioned tasks, the participants were expected to predict the users into four categories based on the degree of suicide risk, that is "No", "Low", "Moderate", and "Severe". In addition, for each task, the participants were expected to generate two metrics to distinguish users with suicide ideation (i.e., flagged/not flagged) and users with suicide ideation who require urgent attention (i.e., urgent/not

urgent). Due to the requirements of our proposed architecture and to accommodate our research objectives (i.e., to measure the impact mental disorders have on suicide ideation), we generated flagged/not flagged and the urgent/not urgent metrics within "Task B" and did not make predictions on the degree of suicide risk. The CLPSych 2019 task participants have submitted results using models trained on classical machine learning and deep learning methods. Under classical machine learning approaches, many have used logistic regression and Support vector machine (SVM) algorithms with manually engineered features, while convolutional neural networks were widely used in deep learning. The CLPSych 2021 shared task can be considered the first of its kind to bring the researchers to the data rather than taking it to the researchers. Given the sensitive nature of the data, the shared task organizers have created a secured data enclave where researchers can work with sensitive data. The data for the task is provided by Qntfy, which is collected through the online platform OurDataHelps.org. The online platform is for individuals to donate data that will be used for mental health research. The shared task consisted of two subtasks where one predicts the suicide risk before the date of an attempt given 30 days of tweets. The second subtask predicts suicide risk given six months of tweets before the attempt or a corresponding date. In total, there were 3,631 users and out of which 1,613 have attempted suicide. From the 1,613 users, the task organizers have only selected users who have completed the questionnaire with birthdate and gender. The task participants have used both classical machine learning and deep learning methods in order to predict at-risk users. For subtask 1, Bayram and Benhiba (2021) have implemented a weighted ensemble approach that combined four machine learning methods: logistic regression, two Naïve Bayes methods and a linear Support Vector Machine. Their approach produced the

best results only for subtask 1. For subtask 2, Gamoran et al. (2021) produced the best results using a Bayesian model with manually engineered features.

An overview of the literature in detecting suicide ideation using social media data is mentioned in Table 2.2. The table identifies the individual research, its objective, the different features and the methods used in generating those features and the main algorithm or the approach used in determining individuals with suicide ideation.

Mentioned below are the abbreviations used in Table 2.2: CLPSych (Computational Linguistics and Clinical Psychology), KNN (K-Nearest Neighbors), CNN (Convolutional Neural Network), RNN (Recurrent Neural Network), LSTM (Long Short-Term Memory), GRU (Gated Recurrent Unit), BERT (Bidirectional Encoder Representations from Transformers), GLoVe (Global Vectors for Word Representation).

Literature	**Objective**	**Features**	**Methods**
Benton, Mitchell, and Hovy (2017)	Mental health including suicide ideation (Twitter data)	Character n-grams	Multi-task learning (i.e., using feedforward neural networks)
Nobles et al. (2018)	Suicide risk assessment (personal communication data, social media data, web browsing history, mental health history)	LIWC psycholinguistic features, word occurrence (TF-IDF)	Feedforward neural networks
Coppersmith et al. (2018)	Suicide risk assessment (social media data)	Neural features (GloVe)	LSTM and self-attention

Literature	Objective	Features	Methods
Matero et al. (2019) (CLPSych 2019)	Suicide risk assessment (Reddit data)	Neural features (BERT), open vocabulary, affect, intensity, NRC lexicon scores	LSTM, GRU, RNN
Mohammadi et al. (2019) (CLPSych 2019)	Suicide risk assessment (Reddit data)	Neural features (GLoVe, ELMo)	A fusion approach using LSTM, GRU, RNN, CNN and SVM models
Ambalavanan et al. (2019) (CLPSych 2019)	Suicide risk assessment (Reddit data)	Neural features (BERT)	BERT
Ruiz et al. (2019) (CLPSych 2019)	Suicide risk assessment (Reddit data)	Clinical findings, semantic role labelling, forum posting behaviour, sentiment	SVM, Naïve Bayes, ensemble model
Chen et al. (2019) (CLPSych 2019)	Suicide risk assessment (Reddit data)	User behavioural features (e.g., posting behaviour, sentiments, motivation, posting content)	SVM (i.e., behavioural model), language model (suicide language model), a hybrid model (combining the behavioural model and the suicide language model)

Literature	Objective	Features	Methods
Morales et al. (2019) (CLPSych 2019)	Suicide risk assessment (Reddit data)	TF-IDF, LDA topics, personality features, tone features, Neural features (FastText, Skip-gram)	Neural Network Synthesis (Ne- uNetS), CNN, LSTM, Random Forest
Bitew et al. (2019) (CLPSych 2019)	Suicide risk assessment (Reddit data)	TF-IDF, emotion, suicide risk	SVM, logistic regression, ensemble model
Gaur et al. (2019)	Suicide risk assessment (Reddit data)	Neural features (ConceptNet embeddings), characteristic features (e.g., lexical, syntactic, sentiment, emotion, mood, upvotes, downvotes)	CNN, SVM, feedforward neural network
Tadesse et al. (2019)	Suicide ideation detection (Reddit data)	Neural features (Word2vec)	LSTM, CNN
Ji et al. (2021)	Suicide ideation detection (Reddit and Twitter data)	lexicon-based sentiment, LDA topics	LSTM, relation networks with attention
Morales et al. (2021) (CLPSych 2021)	Suicide risk assessment (Twitter data)	Grammatical features, character TF-IDF features	Gradient boosted classifiers, ensemble voting classifier (logistic regression, multinomial naïve bayes, random forest), BERTweet model

Literature	Objective	Features	Methods
Wang et al. (2021) (CLPSych 2021)	Suicide risk assessment (Twitter data)	Part-of-speech tags, suicidal behaviour related features, emotion, latent features (Doc2vec)	C-Attention network, SVM, KNN

Table 2.2: Suicide ideation detection related literature summary

According to table 2.2, we could see that many researchers have adopted deep learning methods instead of classical machine learning approaches. Even though many have used the deep learning methods, we could still identify a considerable number of research that have used classical machine learning algorithms such as SVM and logistic regression. In CLPSych 2021, many have used classical machine learning methods instead of neural network architectures. One of the key reasons to use classical machine learning methods could be the dataset being small (Morales et al., 2021) and training a deep neural network with limited data could make the model overfit and not generalize well on the unseen data. On the contrary, the best results in the CLPSych 2019 shared task are obtained using deep neural network architectures. For example, Matero et al. (2019) and Mohammadi et al. (2019) have used RNN based architectures to get the best performing models. Since we are comparing our models with CLPSych 2019 submissions, the best performing models ranked up to fifth place are mentioned in table 2.2. Even though the best performing models (i.e., for Task A, B and C) were trained using the RNN networks, our best performing models are based on the CNN architecture used to extract shared and task-specific hidden features for suicide ideation and mental illness detection. Considering the literature that uses deep neural network architectures, we could not identify any that have used multi-task learning

except for Benton, Mitchell, and Hovy (2017). Unlike the research conducted by Benton, Mitchell, and Hovy (2017), where they have concatenated multiple datasets with further annotations, we have used separate datasets with different distributions without any alterations. Also, the authors have used a feedforward neural network with one shared layer and task-specific layers to predict ten tasks, including seven mental disorders, suicide ideation, neurotypicals, and gender.

One of the key drawbacks of using manual feature engineering is that the impact of certain standard features on the outcome varies between different researches. For example, in contrary to the De Choudhury et al. (2013a) and De Choudhury et al. (2013b), De Choudhury et al. (2014) have identified that features based on emotional measures were less useful in identifying mothers with postpartum depression.

In general, a clear distinction in the lexical and syntactic structure of the language used by individuals with different mental disorders and suicide ideation against neurotypicals can be found throughout the literature. Considering the previous research, we could identify that mental illness and suicide ideation detection have focused mainly on manual feature engineering and classical machine learning algorithms until recent years. The use of neural network models was seen less often, presumably due to the limited amount of annotated data, as deep neural networks thrive on more data than traditional classification algorithms. Also, the neural networks provide less explanation about the features that contribute to the model's performance. Empirically, the requirement of discovering the most prominent features to identify the signs of mental disorders or suicide ideation can be classified as less critical in specific applications. For example, if used in a mental health support forum, it could be argued that identifying the users who require immediate attention is the

primary concern rather than identifying the features that derived the conclusion. If a neural network model produces better predictions than a model trained on manually engineered features, the model trained using the neural network architecture can be used as the predictive model, assuming that the model generalizes well on unseen data. Considering the previous and current research trends in mental illness and suicide ideation detection, we could identify that using neural network models in situations where it is permissible could be more intuitive than using classical machine learning methods with manually engineered features. The proposed research in mental illness and suicide ideation detection is based on deep learning architectures, specifically multi-task learning models, but will consider emotion as an auxiliary input. Intuitively, we can recognize emotion to be strongly associated with mental illnesses and suicide ideation, and this has proven to be the case when evaluating the related literature where many research projects (i.e., on either mental illness or suicide ideation detection) have successfully used emotion as a manually engineered feature with classical machine learning methods.

Chapter 3

Datasets

This chapter presents the datasets we have used to research mental illness and suicide ideation detection and their impact on each other. Details about each dataset are mentioned below, with the related information about the task in which the data was used.

3.1 Research Datasets

3.1.1 Twitter Emotion Classification Dataset

We use the data from the 8th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA-2017). The data was used in the shared task to identify emotion intensity (Mohammad & Bravo-Marquez, 2017). The dataset is used to train the model to predict the emotion category of the tweets used by the users in the CLPSych 2015 Twitter dataset. The emotion categories are only used with the "mental illness detection" experiments with the proposed multi-task learning approach and were not used in our research on "suicide ideation and mental illness detection". We can use these categories in our future work to discover the

possibilities of enhancing mental illness and suicide ideation prediction in a cross-platform knowledge transferring environment. We can use the emotion classification model only to predict the emotion categories of tweets, given that the Reddit posts contain more characters and are comparatively more structured than individual tweets. The dataset is not used in combination with the CLPSych 2015 data but is only used to train a separate model to predict the emotion categories of the tweets within the CLPSych 2015 dataset. The tweets in the dataset were assigned with the labels: anger, fear, joy, sadness, and their associated intensities. Table 3.1 presents the detailed statistics of the dataset.

Emotion	**Train**	**Test**	**Dev**	**Total**
Anger	857	760	84	1,701
Fear	1,147	995	110	2,252
Joy	823	714	79	1,616
Sadness	786	673	74	1,533
Total	3,613	3,142	347	7,102

Table 3.1: WASSA-2017 emotion classification data

The dataset contains 194 tweets that belong to multiple emotion categories. For example, the tweet: "I feel like I am drowning. #depression #anxiety #falure #worthless" is associated with the label's "fear" and "sadness". The authors collected tweets using 50 to 100 query terms that are in relation to the emotion categories. The tweets collected span across three weeks starting from 22nd November 2016. The final dataset consists of tweets that were queried in three different ways: The presence of the query term in the hashtags at the end of the tweet, the same set of tweets as before but without the hashtags that contained the query term and finally, the query term as a word within the tweet or as a hashtag within tweet but not at the end of the tweet. Finally, the master dataset was annotated to include the emotion intensity.

Due to the low amount of data, we used both the training and the test datasets as the training data and tested the trained model on the development dataset. Training on more data could reduce the negative impact of overfitting while increasing the model generalizability. We did not consider the emotional intensity during our training but will consider using it in our future work.

3.1.2 Twitter Mental Illness Detection Dataset

To detect multiple mental disorders, we used the dataset from the Computational Linguistics and Clinical Psychology (CLPsych) 2015 shared task (Coppersmith, Dredze, Harman, Kristy, & Mitchell, 2015), which contain Twitter users that were labelled as being diagnosed (i.e., self-reported diagnoses) with depression, Posttraumatic Stress Disorder (PTSD), and not having either one of the mental disorders (i.e., the control group). Table 3.2 presents the detailed statistics of the dataset.

	Control	**PTSD**	**Depression**
Number of users	572	246	327
Average age	24.4	27.9	21.7
Gender (female) distribution per class	74%	67%	80%

Table 3.2: CLPSych 2015 shared task dataset statistics

The task organizers have taken the same approach as Coppersmith et al. (2014b) to collect the public tweets to identify users susceptible to mental disorders. The dataset contains public tweets identified using their diagnostic statements. The most recent 3,200 tweets from each user were collected, and half of the tweets that mentioned the diagnosis (i.e., depression only) were analyzed to identify the genuineness of the tweets. The tweets indicative of the diagnosis were then removed to reduce the bias on the learning algorithms. Further, the users that did not have at least 25 tweets

were removed, and only the English tweets were retained. Due to the impact that age and gender have on mental disorders, task organizers have predicted the age and gender of the collected users by analyzing their tweets.

We have used the CLPSych 2015 dataset as a standalone dataset when detecting multiple mental disorders (i.e., depression and PTSD) using multi-task learning. In addition, we used the same dataset to identify the impact mental disorders have on suicide ideation detection (chapter 6). Even though the datasets are from two different social media platforms (i.e., mental illness detection data from Twitter and suicide ideation detection data from Reddit), the objective was to investigate if knowledge can be shared among the tasks when detecting suicide ideation and mental disorders in the proposed multi-task learning environment.

3.1.3 The University of Maryland Reddit Suicidality Dataset

The University of Maryland Reddit Suicidality dataset (hereafter known as the UMD dataset) (Shing et al., 2018) is a collection of users who had published posts in the SuicideWatch subreddit. In addition to the posts published in the subreddit, the authors have collected all the posts published in Reddit by the selected users and have filtered those who have posted less than ten posts in total. We used a subset of the UMD dataset for our experiments, which was made available through the CLPSych 2019 shared task on "Predicting the degree of suicide risk in Reddit posts" (Zirikly et al., 2019). Unlike the dataset released by Shing et al. (2018), the CLPSych 2019 dataset does not contain expert annotations but only crowdsource annotations. Even though the annotations provided by the experts are expected to provide more accurate insights into the user's suicide ideation, the task organizers have released

only the crowdsourced annotations due to two main reasons. Unlike the complete dataset, where the expert annotations are used as the test data, the task organizers have provided crowdsource data for training and testing to avoid mismatches when used with machine learning algorithms. The second reason not to release the expert annotated data for the shared task is to make the data available for future shared tasks so that the results obtained using the different kinds of annotators can be compared. Even though we had access to both the expert annotated dataset and the crowdsourced data, we evaluated our trained models mainly on the test dataset released to the CLPSych 2019 shared task participants. We used the crowdsourced data so that we could compare our results against the state-of-the-art research. In addition, we evaluated our models on the data annotated by the experts to see to what extent the models trained using our proposed architecture can be generalized. For future work, we will be using the expert annotated data in combination with the crowdsourced data for training to explore its impact on the model performances when predicting users with suicide ideation.

Shing et al. (2018) have filtered 934 users for annotation out of 11,129 who have posted on the SuicideWatch subreddit. All the posts published by the filtered users were collected irrespective of the subreddit where the posts were published. A matching control group was created from Reddit users where none of the control users have published their posts in any mental health-related subreddits. For expert annotations, 245 users were selected on random to be categorized into four risk categories which are: "No Risk", "Low Risk", "Moderate Risk" and "Severe Risk". The crowdsource workers have annotated 865 users, with 0.554 Krippendorff's alpha inter-annotator agreement.

Table 3.3 states the details about the dataset with the number of users for each class and indicates whether it is annotated by an expert or by crowdsourcing.

Annotator	Number of users					
	No Risk	**Low Risk**	**Moderate Risk**	**Severe Risk**	**Control**	**Total**
Crowdsource	159	63	141	258	621	1,242
Expert	36	50	115	44	245	490
Total	195	113	256	302	866	1,732

Table 3.3: UMD Reddit Suicidality dataset

Table 3.4 states the training and test dataset details for the CLPSych 2019 shared task.

Annotator	Number of users				
	No Risk	**Low Risk**	**Moderate Risk**	**Severe Risk**	**Total**
Crowdsource Train	127	50	113	206	496
Crowdsource Test	32	13	28	52	125
Total	159	63	141	258	621

Table 3.4: CLPSych 2019 crowdsourced dataset details

Even though the inter-annotator agreement for the crowdsourced data is low, the disagreement between the expert annotations and the crowdsourced annotations is less given the flagged/not flagged and urgent/not urgent tasks. According to table 3.5, which represents 245 users annotated by both crowdsourced and expert annotators, Shing et al. (2018) has identified that the disagreement between the two groups of annotators is due to misclassifying low-risk users into the higher-risk categories.

According to the confusion matrix (i.e., table 3.5), which lists crowdsourced annotations against the expert annotations, we could identify an F1-score (i.e., for the positive class) of 0.9385 and 0.8458 for the tasks flagged/not flagged and urgent/not urgent, respectively. Considering these scores, we could identify strong inter-annotator reliability between the expert and crowdsourced annotations. Even though we trained

Expert annotations	Crowdsourced annotations			
	No Risk	**Low Risk**	**Moderate Risk**	**Severe Risk**
No Risk	29	1	1	5
Low Risk	11	13	20	6
Moderate Risk	6	11	47	51
Severe Risk	1	1	8	34

Table 3.5: Crowdsourced against expert annotations

our models using only the crowdsourced data, we also used the expert annotated dataset to demonstrate how well the trained model has generalized.

The CLPSych 2019 shared task focused on three subtasks:

Task A: To predict the level of suicide risk using only the posts published in the SuicideWatch subreddit. Due to the reason that the task provides a minimum amount of data and the approach taken to prove our objective uses deep learning methods that require a sufficient amount of data, we did not conduct any experiments related to "task A".

Task B: All our experiments are based on "task B", where we try to predict the level of suicide risk by taking into account all the posts published in Reddit by the filtered SuicideWatch subreddit users. The task provided us with an ample amount of data so that we could conduct our research.

Task C: The key objective of the task is to identify the level of risk before users publish any content in the SuicideWatch subreddit. None of the posts published in the SuicideWatch by the filtered users are considered. Because our goal is to identify mental disorders' impact on suicide ideation, we did not consider the particular task.

Even though "task B" is to predict the level of risk, our main objective is to identify the possibilities of extracting a shared feature space between the users with suicidal thoughts and mental disorders to improve the predictability of users with a mental

disorder or suicide ideation. To prove our hypotheses, we selected two subtasks from "task B," which is to distinguish users with suicide ideation from the ones that do not have suicidal thoughts (i.e., classification task of flagged/not flagged) and to distinguish the users with suicide ideation that requires urgent attention from the ones that does not (i.e., classification task of urgent/not urgent). For the binary classification task "flagged/not flagged", the users annotated as having "Low," "Moderate", and "Severe" risks' were combined to form the positive class while the users with "No" risk were considered as the control group. We randomly selected users from the table 3.3 control group to merge with the "No" risk users to accommodate our architectural requirements. The number of users selected from the control group (i.e., from 621 control users) is the difference between the positive class and the number of "No" risk users. Similarly, for the binary classification task of predicting urgent/not urgent users, the users belonging to the "Moderate" risk and "Severe" risk categories were combined to form the positive class while the users belonging to the "No" risk and "Low" risk groups were combined to form the control group. To accommodate the difference between the positive and control groups, we randomly selected users from the table 3.3 control group. A detailed distribution of the classes is mentioned in table 3.6. In addition to the crowdsourced data, table 3.6 contains the expert annotated data statistics used at inference to prove further the generalizability of the trained model using the proposed multi-task learning architecture.

Even though we balanced the positive and the negative class distributions by randomly selecting control users, we did not balance the test dataset for a fair comparison with the state-of-the-art results published by the other researchers. The decision not to balance the test data could negatively impact the model performances

Class label	Flagged / not flagged			Urgent / not urgent		
	Train	**Test**	**Test (expert)**	**Train**	**Test**	**Test (expert)**
No risk	127	32	36	127	32	36
Low risk	50 (+)	13 (+)	50 (+)	50	13	50
Moderate risk	113 (+)	28 (+)	115 (+)	113 (+)	28 (+)	115 (+)
Severe risk	206 (+)	52 (+)	44 (+)	206 (+)	52 (+)	44 (+)
Random control	242	-	-	142	-	-
Total	738	125	245	638	125	245

Table 3.6: UMD Reddit suicidality dataset with detailed class distributions. The (+) sign indicates the positive class.

where the class distribution between the shared task training and test data is balanced accordingly.

3.1.4 Self-Reported Mental Health Diagnoses dataset

The Self-reported Mental Health Diagnoses dataset (hereafter known as the SMHD dataset) (Cohan et al., 2018) is used for one of the proposed multi-task learning architecture tasks to detect a mental disorder while sharing certain hidden features with the suicide ideation detection task. The dataset consists of nine common mental illnesses and the control group. The nine mental disorders include Autism, Attention-Deficit Hyperactivity Disorder (ADHD), Post-traumatic Stress Disorder (PTSD), Obsessive-Compulsive Disorder (OCD), Bipolar Disorder, Schizophrenia, Eating Disorder, Anxiety and Depression. A single user could have reported either one or more mental disorders, where a maximum of six mental disorders was identified from specific users. The dataset contains predefined splits for training, development and test data that the authors have used in their preliminary experiments. Table 3.7 presents the number of users under each mental disorder for each data partition.

From all three partitions, it could be identified that more than 90% of the users are

	Number of users			
Disorder	**Train**	**Development**	**Test**	**Total**
Autism	479	480	517	1,476
ADHD	1,768	1,747	1,779	5,294
PTSD	528	516	558	1,602
OCD	409	477	390	1,276
Bipolar	1,216	1,182	1,247	3,645
Schizophrenia	238	278	267	783
Eating	104	115	112	331
Anxiety	1,711	1,593	1,675	4,979
Depression	2,662	2,574	2,611	7,847
Control	92,725	92,420	94,415	279,560
Total	101,840	101,382	103,571	306,793

Table 3.7: SMHD dataset with the number of users under each mental disorder and data partition.

from the control group, followed by users susceptible to having depression, ADHD and anxiety. The minority classes include eating and schizophrenia disorders, representing 0.1% and 0.2%, respectively. Certain users in the dataset were identified with more than one mental disorder, where the majority of such users were identified as having both "anxiety and depression" followed by "bipolar and depression" and "ADHD and depression". For users who self-declared three mental disorders, the most common disorders to be identified were "ADHD, anxiety and depression". Figures 3.1 (from training data), 3.2 (from validation data), and 3.3 (from test data) present several examples of users who self-declared a single or multiple mental disorders.

Unlike the counts from table 3.7, figures 3.1 to 3.3 do not include the same user in more than one mental disorder where only a single user will be represented within each mental disorder or a combination of mental disorders.

When collecting the dataset, authors have used precise patterns to identify users who have self-reported mental illness and users who have used terms related to the researched mental disorders. The keywords to be used in the search were taken from

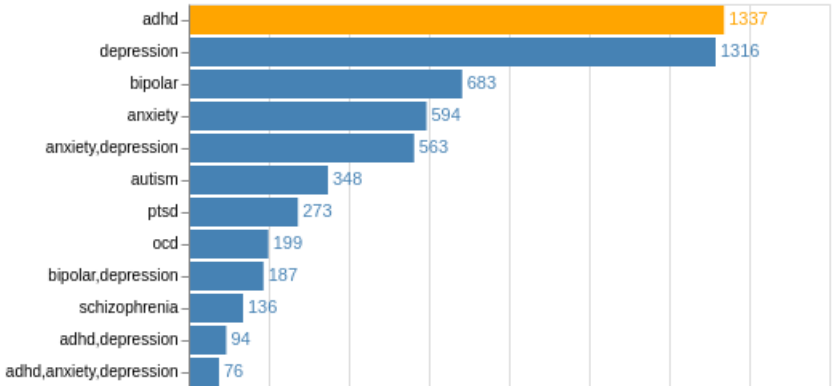


Figure 3.1: SMHD train: class distribution for single or multiple mental disorders.

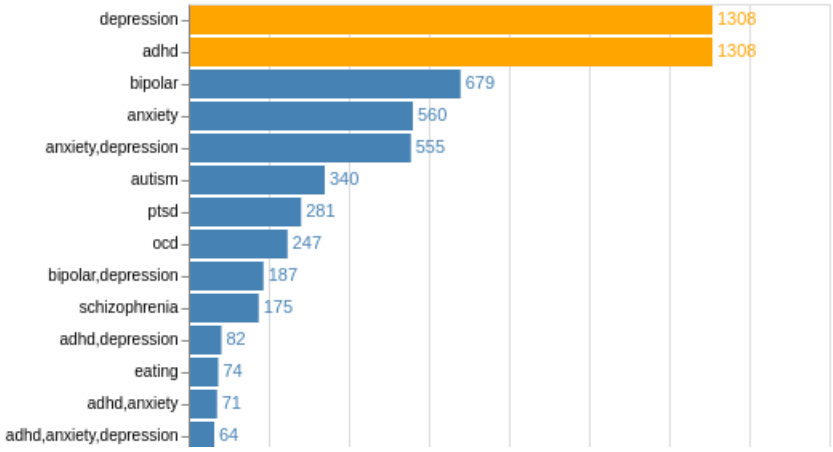


Figure 3.2: SMHD development: class distribution for single or multiple mental disorders

the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). The users whose posted text matched the search terms were selected, and their posts published within, and including January 2006 and December 2017 were added to the overall dataset. The posts that contained the search terms were removed to eliminate the bias such terms could have on the learning algorithm. From all the self-reported posts that indicated diagnosis from nine mental disorders, 500 posts were randomly selected for annotation and from which keywords were extracted to improve the list of terms used

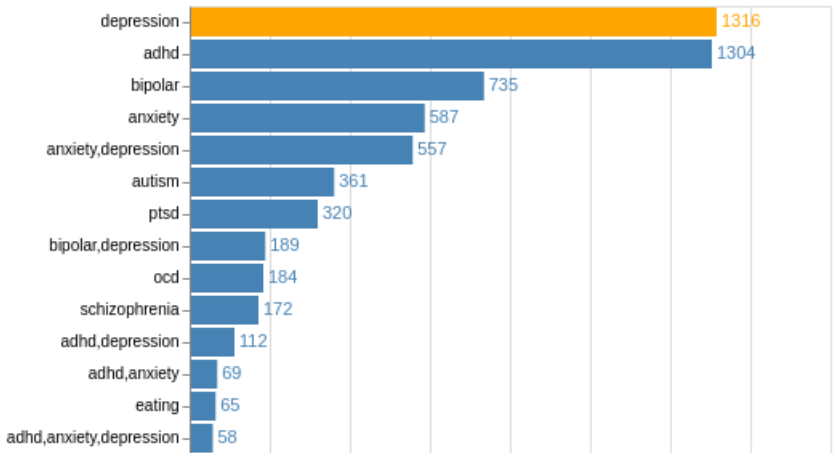


Figure 3.3: SMHD test: class distribution for single or multiple mental disorders

to build the search patterns.

We did not use the entire datasets mentioned above for our research, but only a random sample from the combined train, validation and test datasets. Because we have used two different datasets (i.e., UMD and SMHD) created by two research groups for two different tasks (i.e., for suicide ideation detection and mental illness detection), we had to reorganize the datasets so that the task will be aligned (even though we do not have the same users in the two tasks). The reorganization steps of the datasets were implemented according to the proposed multi-task learning with mixed-parameter sharing architectural requirements and to accommodate the primary research objective, which is to identify the impact mental disorders have on suicide ideation detection.

After pre-processing the data, first, we selected users by filtering them according to the number of tokens used. The filtering identifies users with the number of tokens between the minimum and maximum tokens identified from the pre-processed and concatenated posts from the UMD suicidality dataset. The particular step was

taken to enhance the model training efficacy where sparse matrices due to padding could reduce the training efficiency, with a larger memory footprint and processing requirements. For example, after pre-processing, the number of maximum tokens per user in the SMHD dataset (i.e., by concatenating train, validation and test datasets) is 156,206 with an average sequence length of 4,556 while in the suicide dataset (i.e., by concatenating train, validation and test) it is 49,964 with an average sequence length of 2,602. For each data partition (i.e., training, validation and test) in the SMHD dataset, we calculated the number of mental disorders each user self-declared to identify the level of impact single and multiple mental disorders have on suicide ideation.

Each task in the multi-task learning architecture comprises binary classification tasks to detect whether the Reddit user has suicide ideation (i.e., flagged/not flagged and urgent/not urgent) or a mental disorder. For each of the experiments, a mental disorder was selected from the eight mental disorders (i.e., without the "eating disorder") mentioned in Table 3.7. Users selected from each data partition are matched with an equal number of control users and concatenated into a single data frame used as the source when selecting a random sample to train and test the multi-task learning model.

The required number of users from the SMHD dataset is based on the number of users provided to the CLPSych 2019 shared task participants. Because we will be comparing our results on suicide ideation detection with the state-of-the-art results provided by the CLPSych 2019 task participants, we used the exact number of samples provided by the CLPSych 2019 shared task organizers.

We did not use users who self-declared "eating disorder" as their primary diagnosis

but considered it as a coexisting disorder. In total, 331 users have self-reported "eating disorder". However, according to the model requirements and to balance against the number of users with suicide ideation (i.e., for the task urgent / not urgent), a minimum of 399 users with a mental disorder must be selected for the mental illness detection task. From 399 users, 319 will be used as the training and validation data, while 80 users will be reserved for testing. The requirement of 399 users is the minimum required to answer one of the two research questions, which is to identify users with suicide ideation needing urgent attention. According to our first research question, which predicts users with suicide ideation (i.e., the task flagged/not flagged), the required number of instances for the positive class is 462, with 369 for training and validation and 93 instances for testing.

The SMHD dataset consists of users who self-reported diagnoses (hereafter referred to as "diagnosed") of single and multiple mental disorders, where selecting a user as an input does not guarantee that the impact mental disorder has on suicide ideation is from a single disorder. Due to this reason, we conducted several experiments whereby selecting users who self-declared a single disorder or users who self-reported a primary and one or more mental disorders. Using the content published by users diagnosed with multiple mental disorders, we could further establish the impact comorbidity of mental disorders have on suicide ideation (Hawton et al., 2013; Nock et al., 2009; Simpson & Jamison, 1999; Zaheer et al., 2020). Table 5.3 (i.e., for the task flagged/not flagged) demonstrates the impact individual mental disorders have on suicide ideation, while Table 5.4 (i.e., also for the task flagged/not flagged) demonstrates the impact a primary and coerced mental disorders have on suicide ideation. Also, the impact suicidal thoughts have on mental disorder prediction can also be identified from both

the tables. The figure 3.4 illustrates the self-reported individual mental disorders and the coexisting disorders.

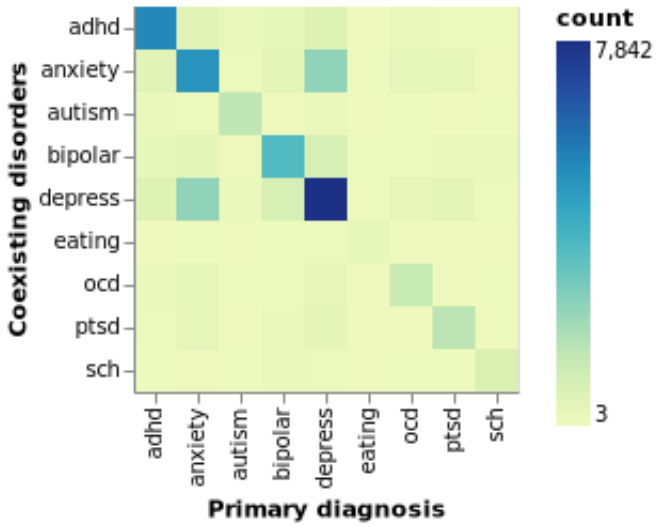


Figure 3.4: Coexisting mental disorders

When training the proposed model to predict users with suicide ideation (i.e., the main objective given the tasks of suicide ideation and mental illness detection within the multi-task learning environment) and who requires urgent attention, we randomly selected 638 users (i.e., 319 with a mental disorder and 319 for the control group) from the SMHD dataset for training and validation, and 125 users for testing. Out of the 125 users, 80 users have self-reported a mental disorder, while 45 have not. The number of users in the test dataset is according to the number of test samples released by the CLPSych 2019 suicide ideation detection "shared task B" and the subtask of detecting users who require urgent attention. Similarly, to predict users with suicide ideation (i.e., the task flagged/not flagged), 738 users (i.e., 369 with a mental disorder and 369 without) were randomly selected for training and validation, while 125 users were selected for testing. Out of the 125 users, 93 users have self-declared one or more

mental disorders, while 32 were not diagnosed with any. Table 3.8 demonstrates the number of users with a single or a multiple mental disorder extracted and concatenated from the SMHD training, validation and test datasets.

Mental disorder	Number of users			
	Single disorder	Multiple disorders	Total	Control
Autism	1,048 (71%)	427 (29%)	1,475	1,475
ADHD	3,944 (75%)	1,344 (25%)	5,288	5,288
PTSD	873 (55%)	726 (45%)	1,599	1,599
OCD	630 (49%)	644 (51%)	1,274	1,274
Bipolar	2,094 (58%)	1,546 (42%)	3,640	3,640
Schizophrenia	480 (62%)	299 (38%)	779	779
Eating	N/A	N/A	N/A	N/A
Anxiety	1,740 (35%)	3,235 (65%)	4,975	4,975
Depression	3,939 (50%)	3,903 (50%)	7,842	7,842

Table 3.8: SMHD combined train, validation and test datasets with users who self-declared a single or multiple mental disorders

The column "Single disorder" states the number of users who have self-reported a single mental disorder, while the column "Multiple disorder" states the number of users with the primary diagnosis and who have self-declared one or more other mental disorders. For example, 1,048 users have self-declared Autism while 427 users have self-declared Autism and one or more other mental disorders. The 427 users were identified with the following disorders:
'PTSD': 22, 'ADHD': 188, 'Eating': 3, 'Depress': 144, Schizophrenia': 28, 'Anxiety': 108, OCD: 46, 'Bipolar': 57

In parenthesis under both single and multiple disorder columns is the percentage of users as a proportion of the total users who have self-declared one or more mental disorders. According to table 3.8, it could be identified that certain users have self-reported coexisting disorders in addition to their primary diagnosis. Even considering figures 3.1 to 3.3, we could identify the density of coexisting disorders.

For example, users whom self-declared ADHD have also declared multiple disorders such as depression and anxiety. Similarly, users who self-declared anxiety, bipolar, ADHD and PTSD could have declared depression as a coexisting disorder.

Even though we did not consider users with the "eating disorder" as a primary mental illness in our research due to its lack of samples, we included the users who self-reported "eating disorder" as a coexisting mental illness.

We conducted our experiments with several random samples taken from the SMHD dataset mentioned in table 3.8 and have reported the best results. The experiments were conducted on five stratified shuffle splits, and for each split, 80% of data was allocated for training and 20% for validation. Taking the best performing random sample is due to the reason that, given the poor results reported by the researchers using the SMHD dataset (Cohan et al., 2018; Harrigian et al., 2020), specific random samples could contain distorted data that could negatively impact the model predictability on both the tasks. Given the proposed architecture, each task in the MTL environment needs to generate better results so that the overall predictability of the model can be enhanced.

3.2 Data Pre-Processing

We used a custom script to remove URLs, @mentions, #hashtags, RTweets, emoticons, emoji, and numbers. The #hashtags, emoticons and emojis were removed to derive a more generalized vocabulary, but we will conduct further research in future work to discover the impact such features could have on the model's performance. We removed a selected set of stopwords but kept first, second, and third-person pronouns. The first-person singular pronouns are frequently used by individuals with mental disorders

such as depression (Pennebaker et al., 2007). Also, the punctuation marks except for a selected few were removed from the CLPSych 2015 data (i.e., when using MTL to predict multiple mental disorders). The full stops, commas, exclamation points, and question marks were kept while removing the other punctuation marks. However, we removed all the punctuation marks from the Reddit datasets (i.e., UMD and SMHD datasets) and the CLPSych 2015 data when predicting users with suicide ideation and mental disorders. The NLTK tweet tokenizer was used to tokenize the tweets, while the spaCy tokenizer was used to tokenize the Reddit data. We made the text lowercase for all the datasets, expanded the contractions, and removed extra spaces, newline characters, and tabs. The above steps were applied accordingly to all the individual tweets and Reddit posts. We removed any record that returned an empty string after the pre-processing. From the SMHD dataset, we filtered out users who had less than 50 tokens so that the shared feature space would be deep enough to enhance the prediction outcome of the two tasks. Similarly, we removed users who had less than 20 tokens from the CLPSych 2015 dataset. The CLPSych 2015 data filtering was only applied when the data was used for multi-task learning to predict users with suicide ideation and mental disorders within the chapter 6 (Cross-platform knowledge transfer). After pre-processing, the minimum number of tokens used by a user in the UMD dataset is 76.

3.3 Exploratory Analysis

3.3.1 Suicide Ideation and Mental Illness Detection using Reddit Data

For an overview of the datasets being used to detect suicide ideation and mental illness, several analyses were conducted using the Scattertext (Kessler, 2017) that

helps in identifying key characteristics in a given corpus. All the exploratory analyses were conducted using only the train and validation data without including the test data. Figures 3.5, 3.6 and 3.7 give an overview of the different terms used by distinct users identified as having suicidal thoughts (figure 3.5), self-declared PTSD (figure 3.6) and self-declared depression (figure 3.6). The legends of each graph specify the most frequent terms detected from the positive and negative classes. For better visualization, only the rightmost part of the graph is included. We selected users who self-reported PTSD and depression to be studied during the exploratory analysis stage, considering our preliminary experiments. We selected PTSD due to the improved results that we managed to obtain and depression to understand better the comparatively poor performances it produced even though the majority of the individuals who have committed suicide were diagnosed with depression (Hawton et al., 2013). To avoid any bias during inference, we used only the training and validation partitions of the respective datasets for the analysis.

The graphs list terms identified among the users in the positive and control groups where both the "X" and "Y" axis represent the term/phrase frequency in each group. Each blue data point in the graphs represents a word or a phrase used by a user having a mental disorder or suicide ideation. The red points in the graphs represent the control group. The more familiar terms among the users belonging to the positive class are represented in the upper left corner of the graph, while the lower right corner represents terms or phrases more commonly identified among the users in the control group.

In comparison, a clear distinction between the users with suicide ideation and mental disorders can be identified when inspecting the most frequently used terms.

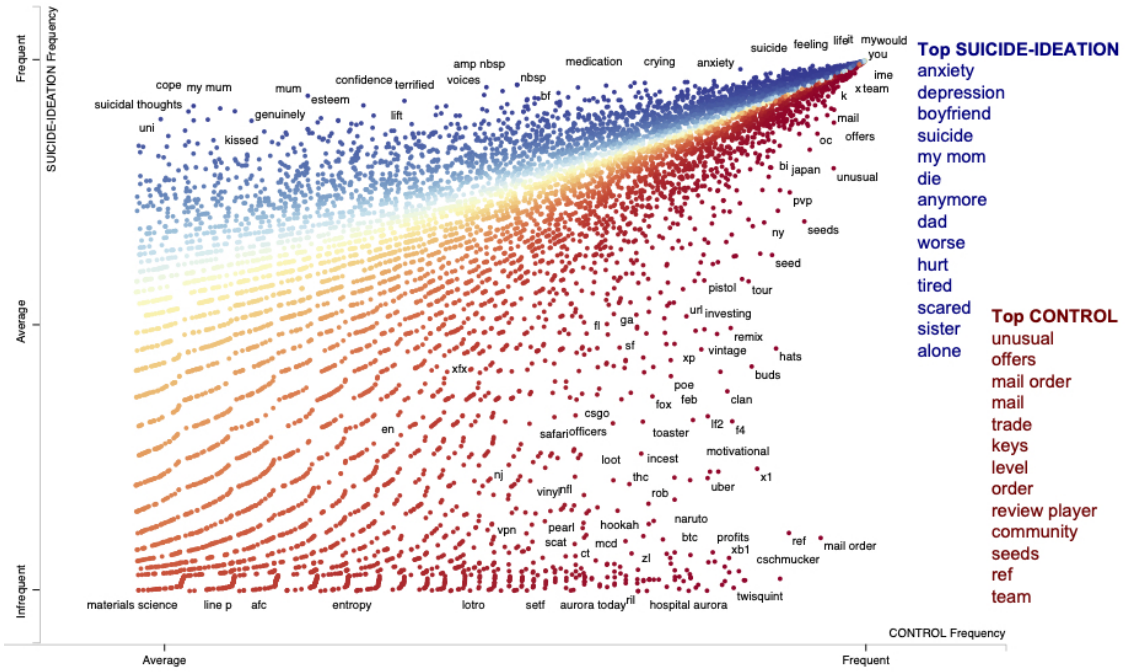


Figure 3.5: The most frequent terms used by individuals with and without suicidal thoughts.

Users have used terms such as "anxiety", "depression", "suicide", "die", "hurt" and "alone", which can be considered as more indicative of the mental state of a person susceptible to suicidal thoughts. According to figures 3.6 and 3.7, a clear distinction between the terms used by the individuals diagnosed with PTSD and depression can be identified. For example, the users diagnosed with PTSD have used terms such as, "abuse", "abusive", "relationship", "therapy", "feelings", and "pain", which could have a strong relationship to traumatic events experienced by the individual (Wilcox et al., 2009). When analyzing the most frequent terms used by individuals who self-declared depression, we could identify that the terms used do not reflect a strong association with depression, which could be because the dataset being used does not reflect too many depressive characteristics. As a result, the model obtained

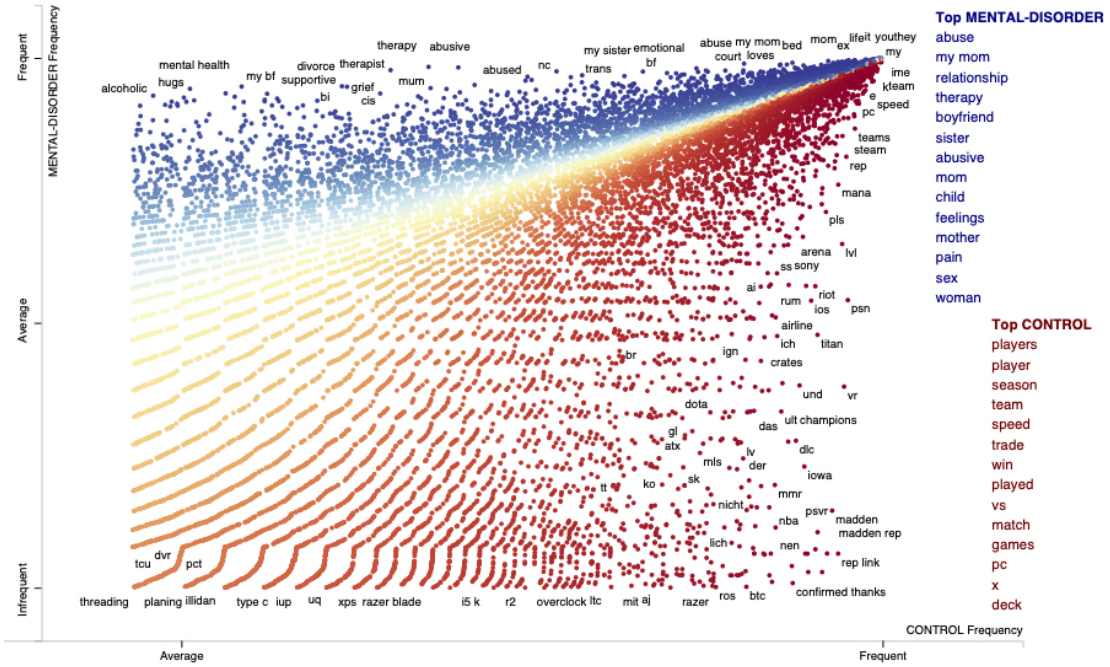


Figure 3.6: The most frequent terms used by users diagnosed with and without PTSD.

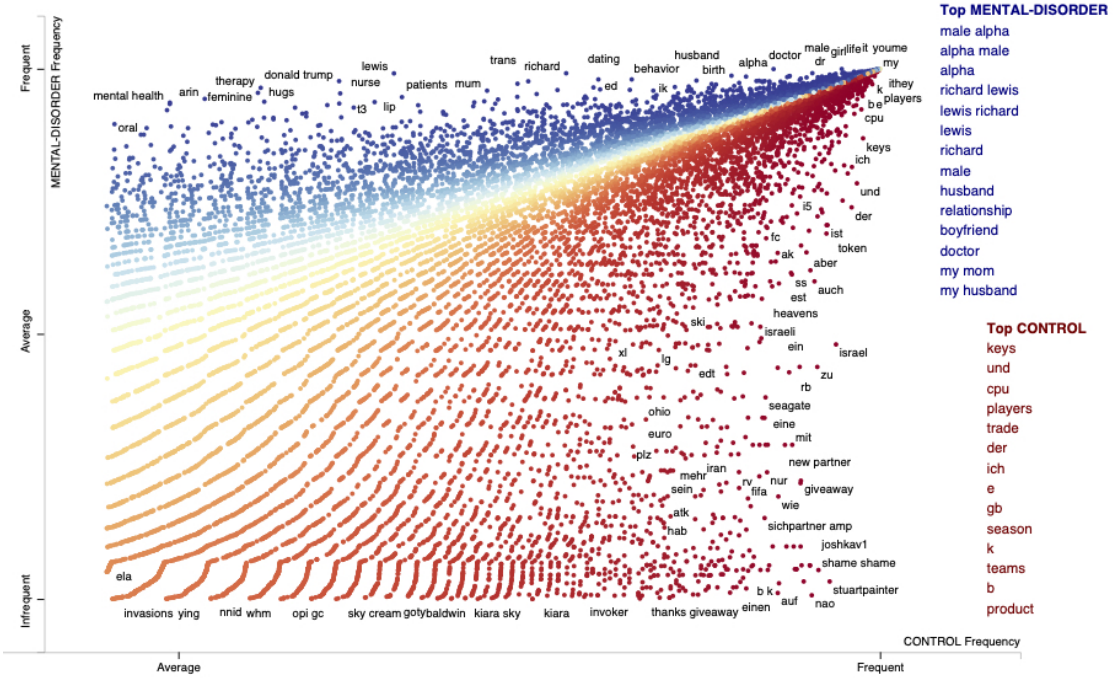


Figure 3.7: The most frequent terms used by individuals with and without depression.

poor performance when used in either multi-task or single-task learning environments. It is important to note that terms with high frequency in a class cannot be used as conclusive evidence to identify a user as having suicidal thoughts or a mental disorder, but only as indicators that can be used as features when training machine learning models.

We used the EMPATH library (Fast et al., 2016) to understand further the emotional and topical categories of the users with suicidal thoughts and mental illnesses. Identifying the unique categories gives an abstract view of the characteristics that can distinguish each user group. Figures 3.8 till 3.10 provide an overview of the EMPATH categories to show how strong each group of users is associated with each category. Figure 3.8 provides an overview of the empath categories identified from users who self-declared only PTSD, while figure 3.9 lists the first fourteen most strongly related EMPATH categories in relation to the terms used by users who self-reported depression. Finally, figure 3.10 lists the categories identified from users having suicidal thoughts. Even though the EMPATH categories are dynamically generated, we used only the most frequent categories as auxiliary inputs.

Based on figures 3.8 to 3.10 and compared to users who self-declared depression, more similarity in EMPATH categories between the users with suicidal thoughts and PTSD can be identified. Users with suicide ideation were collectively identified with the unique categories: 'neglect', 'timidity', 'suffering', 'anger', 'weakness', which were not identified among the users with PTSD and depression.

To discover the likelihood of EMPATH categories being shared among users with mental disorders (i.e., single or multiple) and suicide ideation, first, we merged the users with mental disorders and suicide ideation into the positive class. The

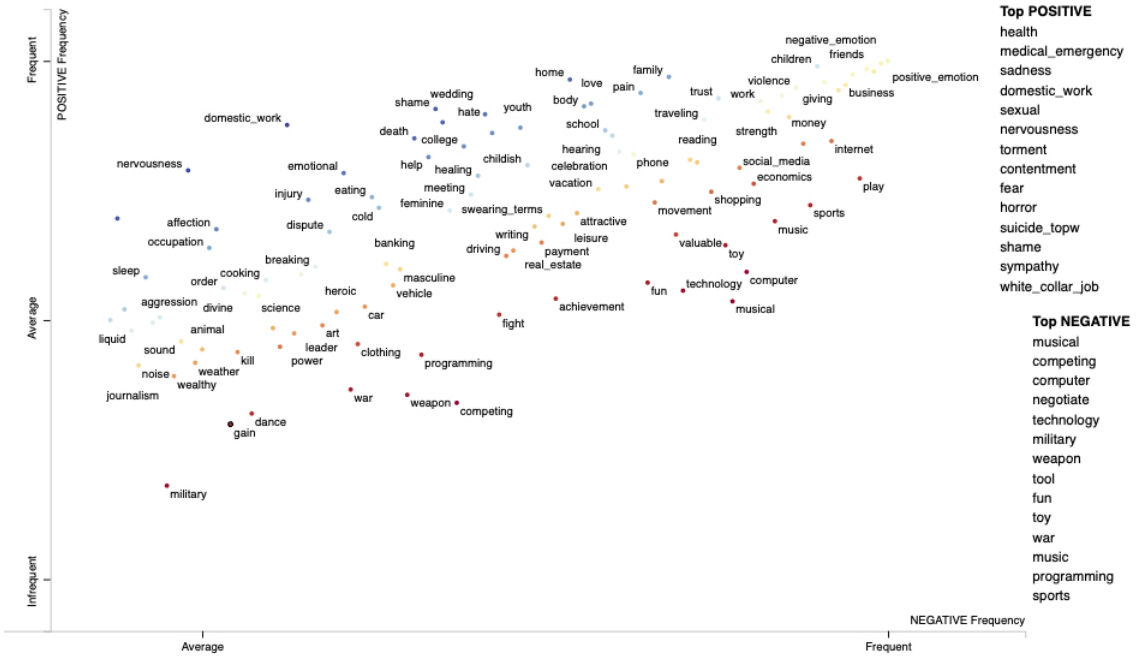


Figure 3.8: EMPATH categories from users diagnosed with and without PTSD.

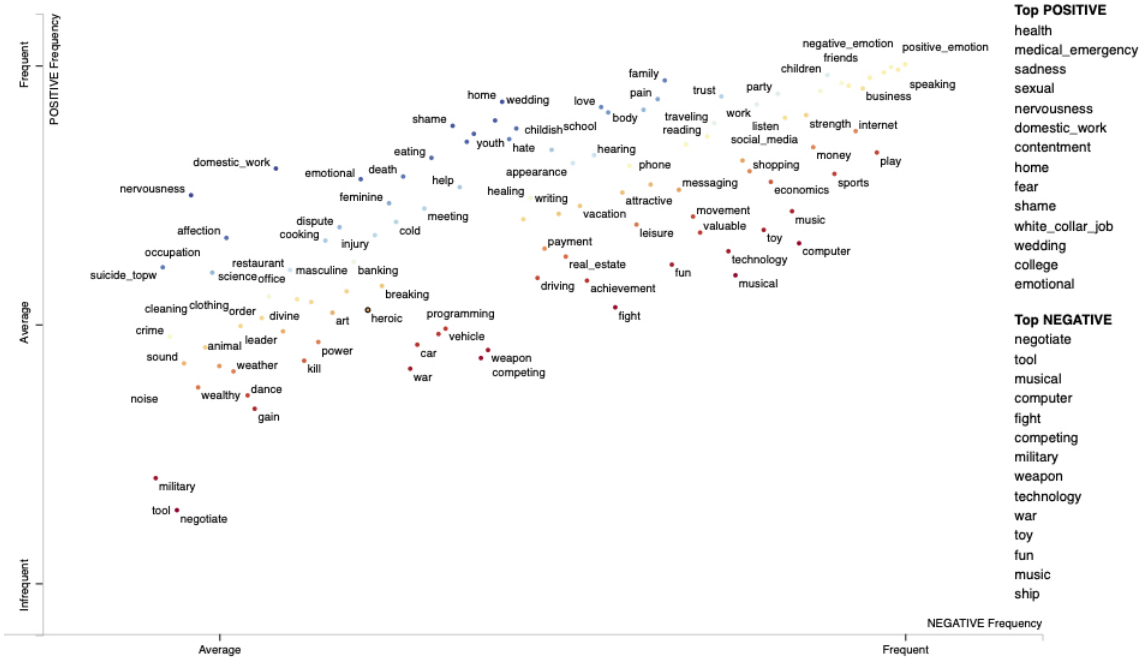


Figure 3.9: EMPATH categories from users with and without depression.

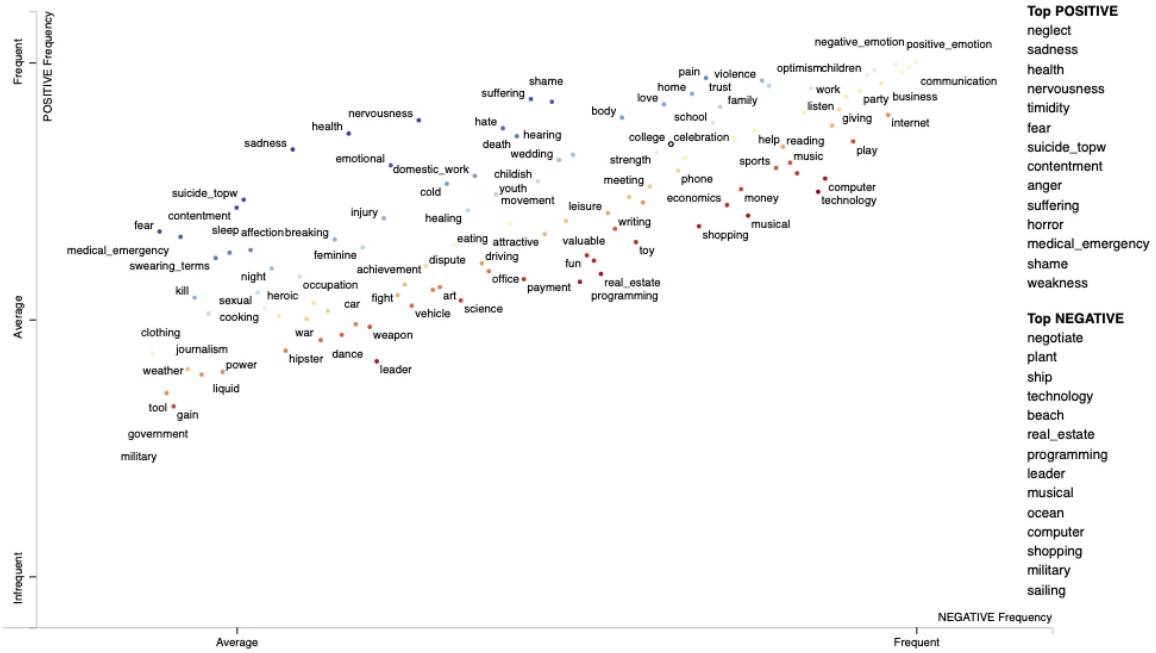


Figure 3.10: EMPATH categories from users with and without suicidal thoughts.

negative class contained users from the control groups. Using the merged datasets, we generated fourteen EMPATH categories for the positive and negative classes. Table 3.9 demonstrates the top fourteen EMPATH categories for different combinations of sampled datasets (i.e., only for users who self-reported PTSD or depression).

According to table 3.9, a considerable overlap between the EMPATH categories among the different permutations of the combined datasets can be identified. The overlap could be because specific terms are more commonly used by users with suicidal thoughts and mental disorders. For example, categories such as "health", "medical_emergency", "sadness", "nervousness", and "fear" can be identified within the content posted by users diagnosed with depression and PTSD as well as among the users with suicide ideation (i.e., according to graphs 3.8, 3.8, and 3.8). In addition to the commonly used categories, certain categories are prioritized differently within each

Suicide ideation and mental disorder	EMPATH Categories (positive class)	EMPATH Categories (control group)
Suicide + Depression (single)	health, nervousness, sadness, masculine, fear, medical_emergency, contentment, power, shame, neglect, lust, timidity, domestic_work, sexual	negotiate, military, technology, weapon, toy, fun, musical, computer, valuable, tool, competing, achievement, ship, economics,
Suicide + Depression (multiple)	health, nervousness, sadness, medical_emergency, fear, contentment, shame, domestic_work, neglect, timidity, suffering, white_collar_job, lust, hate	negotiate, tool, computer, musical, military, technology, toy, competing, fun, weapon, beach, ship, valuable, fight
Suicide + PTSD (single)	health, sadness, nervousness, medical_emergency, contentment, domestic_work, fear, torment, shame, sexual, white_collar_job, horror, neglect, timidity	musical, competing, technology, tool, computer, fun, military, toy, negotiate, weapon, fight, sports, music, ship
Suicide + PTSD (multiple)	health, medical_emergency, sadness, nervousness, fear, contentment, domestic_work, horror, torment, suicide_topw, neglect, shame, suffering, sexual	musical, negotiate, military, technology, competing, tool, fun, computer, weapon, programming, toy, ocean, gain, beach,

Table 3.9: Filtered empath categories from combined datasets

group of users. For example, the category "torment" is ranked higher among the users who self-reported PTSD. Similarly, the term "hate" is ranked higher among users diagnosed with depression. Hypothetically it could be argued that certain mental disorders do share features with users having suicidal thoughts.

Our objective in identifying the EMPATH categories that overlap between mental illnesses and suicide ideation is to investigate the possibilities of using the identified categories as auxiliary inputs to enhance the model performance and its generalizability, and specifically by improving the level of accuracy that differentiates users with suicide

ideation and mental disorders from the control group. A clear distinction can be seen when comparing the EMPATH categories of users with suicide ideation or mental disorders with neurotypical users. Using such distinctive features as auxiliary inputs to a deep learning model could enhance the overall model performance.

In our research, we will be using the common categories between the users with suicide ideation and users whom self-declared PTSD to investigate the possibilities of using such auxiliary features to enhance model predictability and generalization. The reason to use PTSD as the mental disorder to identify the impact EMPATH categories have on the model performances is that our preliminary experiments identified that when using PTSD as the input to the proposed MTL architecture, it produced better performances than other mental disorders.

3.3.2 Suicide Ideation and Mental Illness Detection using Multiplatform Data

To gain an understanding of the use of vocabulary among the Reddit users with suicide ideation (i.e., from the UMD dataset) and Twitter users diagnosed with mental disorders (i.e., users from the CLPSych 2015 dataset diagnosed with either depression or PTSD), we generated the following graphs that represent the most frequently used terms (i.e., figure 3.11) and their associated EMPATH categories (i.e., figure 3.12). The legends of each graph specify the most frequent terms/phrases or the top most strongly associated EMPATH categories identified from the given datasets' positive and negative classes. For better visualization, only the rightmost part of the graphs is included. Through evaluation, it could be identified that when using the Twitter data, the filtered terms/phrases are somewhat unstructured compared to the

terms/phrases used by users from the Reddit social media platform (e.g., figures 3.5, 3.6, 3.7). However, we could still filter terms that could provide valuable insights into differentiating users diagnosed with a mental disorder from neurotypicals. Terms such as "depression" and "anxiety" can be considered features that could get shared between the users with suicide ideation and mental illness.

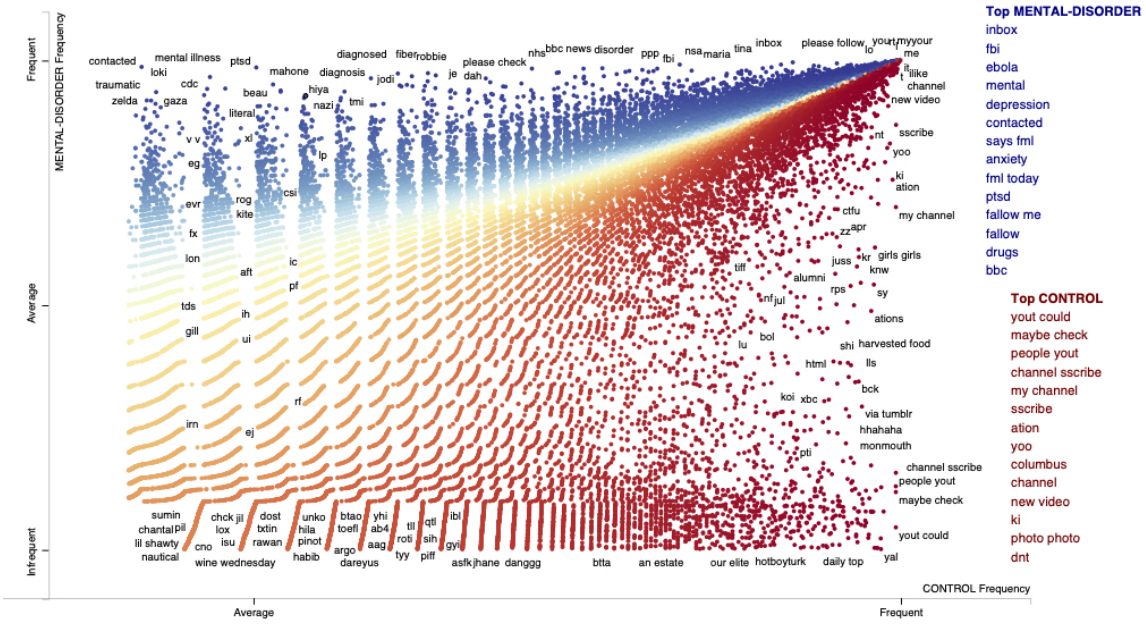


Figure 3.11: The most frequent terms used by individuals with and without mental disorders.

We combined the Twitter dataset with the UMD dataset to discover the shared and ranked EMPATH categories based on the term frequencies. The users with suicidal thoughts (i.e., the positive class) from the UMD dataset were merged with those diagnosed with PTSD or depression from the CLPSych 2015 dataset. The merged control group consists of the users from the control groups of the two respective datasets. According to figure 3.13, we could identify that the most frequent EMPATH categories discovered when using the two datasets alone (i.e., according to figures 3.10 and 3.12)

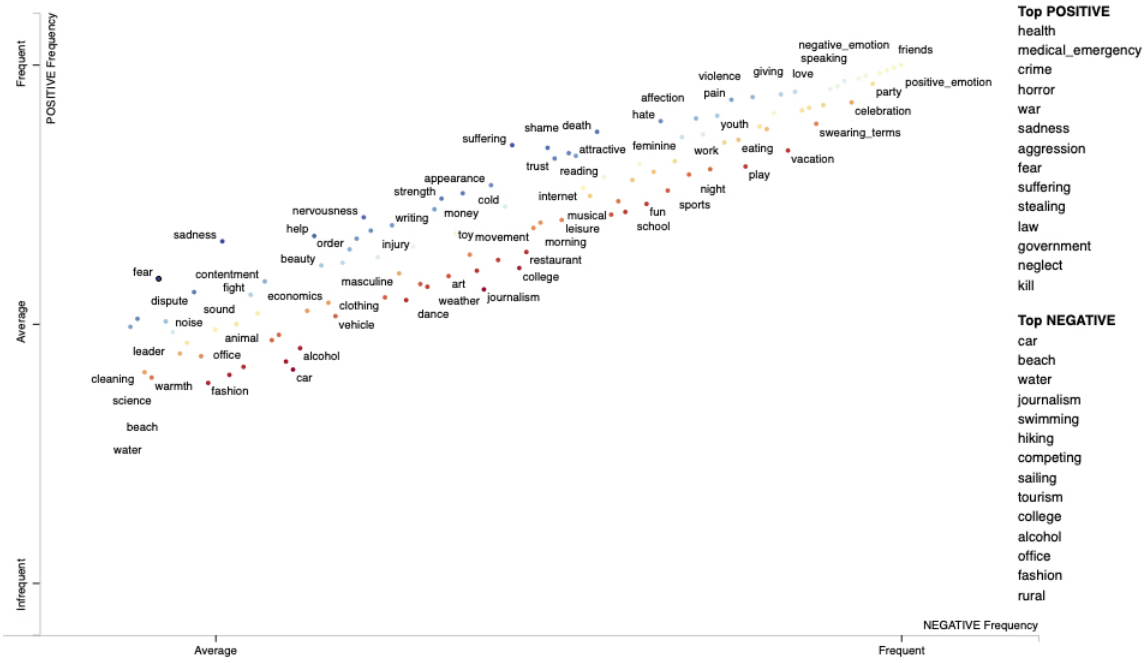


Figure 3.12: EMPATH categories from users diagnosed with and without mental disorders.

are included when using the two datasets as a single merged dataset. Hypothetically, we could argue that using data from two different social media platforms could still discover certain shared features among suicide ideation and mental disorders.

The purpose of using Twitter data in our research is to discover whether knowledge can be shared between tasks that use data from two different platforms. From the exploratory analysis, we could identify that, despite the differences in the vocabularies, users with suicide ideation (i.e., from the Reddit social media platform) share several EMPATH categories with users diagnosed with mental disorders (i.e., from the Twitter social media platform). We conduct experiments using different combinations of the filtered EMPATH categories as mentioned under the "Top POSITIVE" column in figure 3.13 to identify which categories can enhance the model performances.

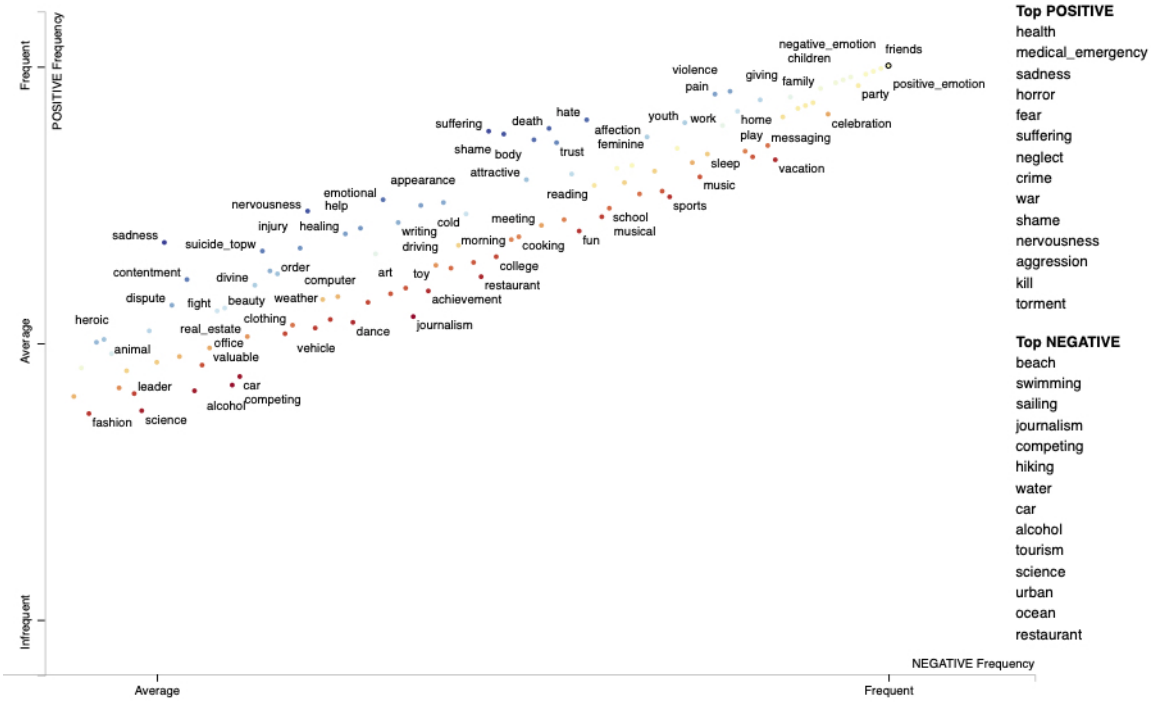


Figure 3.13: EMPATH categories from merged users with suicide ideation and mental disorders.

However, we did not conduct extensive experiments using all the possible combinations because our research objective is to identify the positive impact mental disorders have on detecting individuals with suicide ideation. The primary purpose of using the EMPATH categories as auxiliary inputs is to provide adequate evidence that, given a well-generalized model, we can still use manually engineered features shared between different datasets from different platforms as auxiliary inputs to enhance deep learning model performances.

Chapter 4

Mental Illness Detection using Multi-Task Learning[1]

With the recent success in adopting deep learning methods using social media data to detect mental illnesses (Kshirsagar et al., 2017; Husseini Orabi et al., 2018), we investigate the possibilities of detecting multiple mental disorders (i.e., PTSD and depression) using different deep neural network architectures in comparison to classical machine learning approaches such as the Support Vector Machines. Even with the inherent drawbacks of manual feature engineering, we investigate the adoption of prominent features used with classical machine learning methods as auxiliary inputs to enhance the model's predictability and generalization. These features (i.e., emotion categories) were predicted using a deep learning model trained on data related to different emotion categories (i.e., anger, sadness, joy and fear).

Our proposed solution to detect mental disorders using deep learning methods consists of two key components. The first one identifies the emotion category expressed by each user using the model trained on the WASSA 2017 shared task dataset

[1]Parts of this chapter were published in the paper "Multi-Task, Multi-Channel, Multi-Input Learning for Mental Illness Detection using Social Media Text", by Prasadith Kirinde Gamaarachchige and Diana Inkpen, in the Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019).

(subsection 3.1.1). The second component is a model that predicts users susceptible to PTSD or depression. For both the components, a common base architecture was used. We used a multi-channel Convolutional Neural Network with three different kernel sizes (i.e., 1, 2, and 3) as our base architecture. We recognized that the multi-channel CNN architecture (Y. Kim, 2014) produces better results than the Recurrent Neural Network (RNN) architectures, which can be considered a more appropriate architecture for sequence classification tasks considering their capability in capturing historical information. Also, through extensive experiments, we discovered that using a multi-channel CNN architecture (i.e., with three channels) tends to outperform the single-channel architecture and architectures with two or four channels. It is vital to identify a well-generalized model that does not underfit or overfit the training dataset. A more complex model with limited data could overfit the training dataset, while an uncomplicated model might not discover the function that could effectively map the inputs to the outputs.

For all the experiments conducted, we used the WASSA 2017 dataset and the CLPSych 2015 dataset. Both datasets have been extracted from the Twitter social media platform. To obtain further details about the datasets being used for the following experiments to detect multiple mental disorders using MTL, please refer to section 3.1.1 for the WASSA dataset and section 3.1.2 for the CLPSych 2015 dataset.

4.1 Model Architecture

The selected model architecture consists of three main components: multi-task learning, CNN with multi-channel, and multi-inputs. Multi-task learning was shown to be successful when the data is noisy and limited so that when trying to learn one

task, one could gain additional knowledge from the other tasks to identify the most relevant features. Learning a shared representation so that individual tasks can benefit from one another (Caruana, 1997) can be considered one of the most appropriate architectures when detecting multiple mental illnesses. Benton, Mitchell, and Hovy (2017) demonstrated the successful use of multi-task learning to recognize mental illnesses and suicide ideation. Unlike their approach, we add multiple inputs previously discovered by computational linguistics and psychology researchers to enhance the model performances. We consider that it is vital to identify the impact of the features engineered by researchers on the model's performance. We also recognized that using a CNN multi-channel architecture is best suited for many tasks dealing with limited unstructured data than RNN architectures or multilayer perceptrons (MLP).

4.1.1 Emotion Classification

The neural network architecture used for the multiclass, multi-label classification for emotion is shown in Figure 4.1.

We used the multi-channel model as the base model in emotion classification (i.e., to detect anger, sadness, joy, and fear) and mental illness detection (i.e., PTSD and depression). The multi-channel model uses three versions of a standard CNN architecture with different kernel sizes. We identified that using different kernel sizes (different n-grams sizes) with Global Maximum Pooling produces better results than a standard CNN architecture. The optimal validation accuracy for emotion and mental illness detection models was inferred using three channels with kernel sizes 1, 2, and 3. Increasing the kernel sizes or the number of channels reduced the validation accuracies.

For the emotion classification task, the CNN in each channel was tested with 64

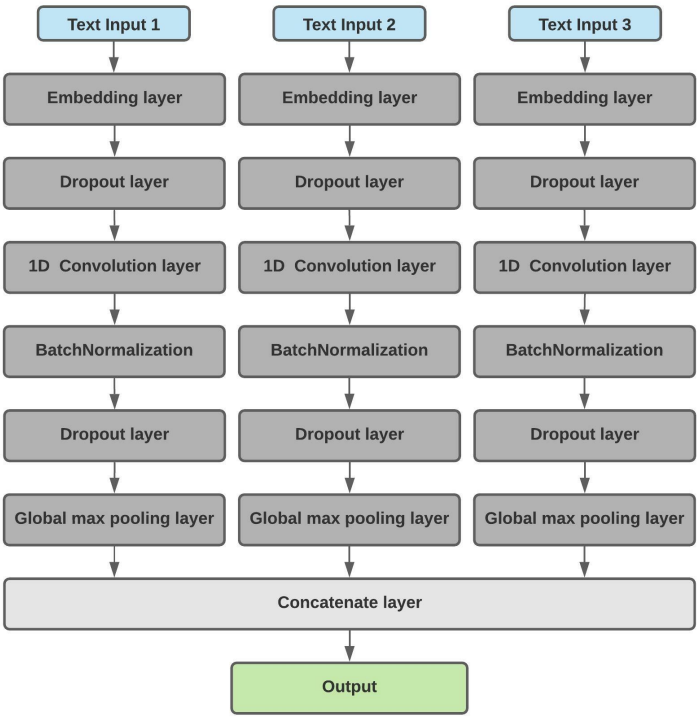


Figure 4.1: Multi-channel convolutional neural network for emotion classification.

filters, same padding and a stride of 1 (distance between successive sliding windows). We used Rectified Linear Unit (ReLU) (Nair & Hinton, 2010) as the activation function. To normalize the data and to reduce the impact of model overfitting, we used batch normalization and dropout (Srivastava et al., 2014) as the regularization method with a probability of 0.2. As the final layer in each channel, we used global maximum pooling to reduce the number of parameters needed to learn so that it could further reduce the impact of model overfitting. The outputs from each global maximum pooling layer (i.e., from each channel) were concatenated and fed into a fully connected layer with four hidden units that use sigmoid activation to generate the output. All the inputs were sent through the trainable embedding layers (randomly initialized)

with a dimension of 300 for the emotion classification task and 100 for the mental illness detection task.

4.1.2 Mental Illness Detection

When building the multi-task learning model to detect mental illnesses, the base architecture (i.e., for the shared representation) has used a structure similar to the one used in emotion classification (i.e., figure 4.1). The fundamental changes to the base model include using 256 filters instead of 64 and using $L1$ kernel regularization in each convolution layer with a regularization factor of 10^{-6}. We used the model trained on the emotion data to predict the emotion category of the individual Twitter messages in the CLPSych 2015 dataset. We grouped the predicted probabilities for each user under the different emotion categories by calculating the standard deviation. As multiple inputs, we used the predicted probabilities for each emotion category when detecting neurotypical and depressed users, while age and gender are used as inputs when predicting users with PTSD. Before concatenating the multiple inputs with the output from the multi-channel architecture, the multiple inputs were transformed using a fully connected layer with 128 hidden units and ReLU activation. The output from the shared layers and the transformed multiple inputs were merged before being used as the input to the output layers with individual hidden units and sigmoid activation. Before applying multiple inputs to the neural network architecture, all the relevant inputs are normalized using a minimum, maximum scaler initialized within the range 0 and 1. The neural network architecture used for the multi-task, multi-channel, multi-input model for mental illness detection is shown in Figure 4.2.

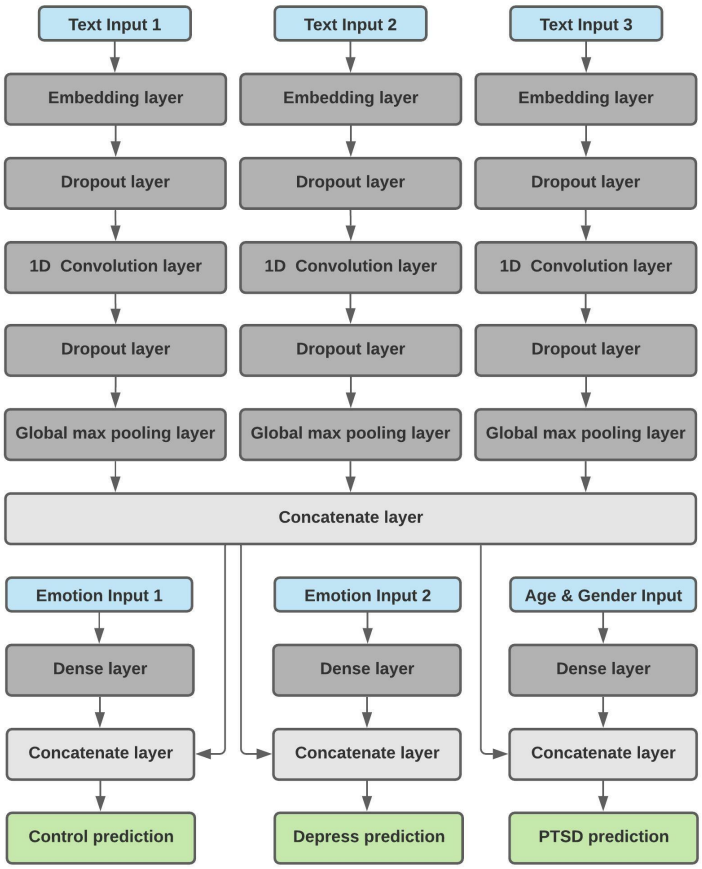


Figure 4.2: Multi-task, multi-channel, multi-input model for mental illness detection

4.2 Experiments

4.2.1 Creating the Vocabulary

After pre-processing the data according to section 3.2, we selected 200,000 unique tokens to build the vocabulary, rather than choosing all the unique words (because the vectors for rare words are not reliable). To obtain an enriched dictionary containing the most relevant terms, we introduced a novel approach instead of the traditional

approach used in Keras deep learning API[2]. Our approach considers the top 'K' terms based on their term frequency and inverse document frequency (TF-IDF) scores. To build the dictionary, we first calculated the TF-IDF values under each user (i.e., by considering all the tweets of a single user as one document). Then we took the maximum score out of all the assigned TF-IDF scores for a given word. The reason for taking the maximum is to extract the words identified as closely related to a given user. Based on the computed values, we rearranged the dictionary in descending order. A dictionary created using the above approach allows the model to capture the underlying relationships between the critical words.

In comparison to the word frequency-based approach, using the vocabulary based on the TF-IDF scores has produced relatively better results for the reported metrics (refer table 4.2). Further, we could also identify a more stable model when using the vocabulary based on the TF-IDF scores (see Figure 4.3). the validation loss (i.e., represented in figure 4.3 (a) y-axis) and validation accuracies (i.e., represented in figure 4.3 (c) y-axis) obtained using the vocabulary generated with words based on their TF-IDF scores demonstrate stable learning with less randomness.

In contrast, the validation loss (i.e., represented in figure 4.3 (b) y-axis) and validation accuracy (i.e., represented in figure 4.3 (d) y-axis), which are based on the vocabulary generated using word frequencies show comparatively unstable learning. The validation loss and the validation accuracy are plotted against the number of epochs (x-axis) and five stratified shuffle splits (noted in the graph's legend as a "fold").

When feeding data to a neural network, all the sequences in a batch must have the same length. Due to this reason, it is essential to select a maximum length for the

[2]https://keras.io/preprocessing/text/

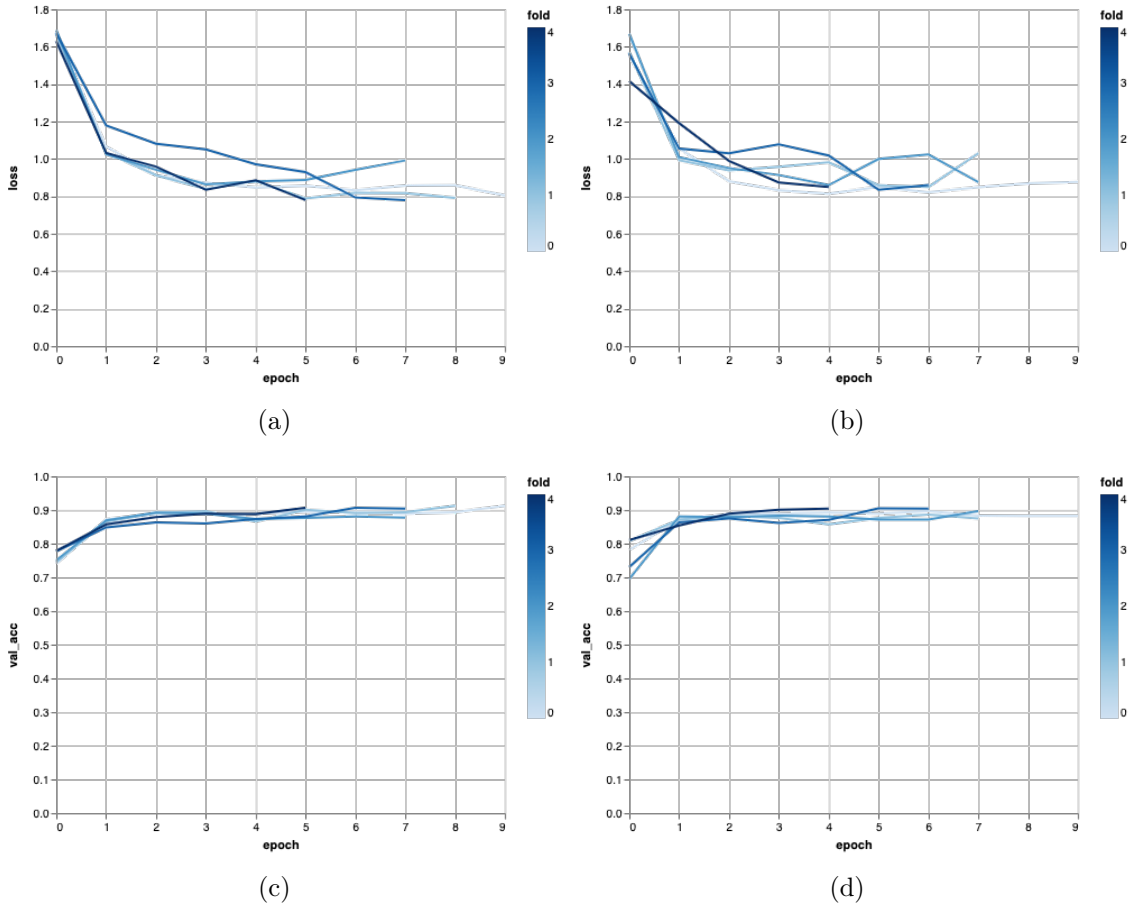


Figure 4.3: ROC curves for: (a) Validation loss when using TF-IDF scores. (b) Validation loss when using word frequencies. (c) Validation accuracy when using TF-IDF scores. (d) Validation accuracy when using word frequencies.

sequence. When choosing the maximum sequence length, it is crucial to capture as much information as possible from each user, especially given the domain of mental illness detection. Since we have concatenated all the individual tweets belonging to one user as a single string, a high variance in the sequence length can be identified among users. On average, a single user has used around 15,800 tokens, whereas the maximum number of tokens used by an individual is around 64,800. Selecting the average

sequence length as the maximum length will make the system have less information to be trained on because there are almost 600 users who have used more tokens than the recorded average. However, if the maximum sequence length is taken as the maximum number of tokens used by a single user, it will generate space and time complexities where sparse matrices will require more memory and extra time to perform the matrix operations. Due to this reason, we calculated the maximum sequence length to cover the sequence length used by the majority of the users. To calculate the new maximum sequence length, we took the value three standard deviations away from the mean sequence length, that is 46,200 tokens covering 99% of users. The shorter sequences were padded with zeros (i.e., to the end of the sequence), and the longer sequences were truncated (i.e., from the end of the sequence).

4.2.2 Model Training

We minimized the validation loss to learn the optimal neural network parameters when training both the emotion and the mental illness detection models. To train both models, we used minibatch gradient descent with smaller batch sizes. Using a smaller batch size is known to stabilize model training while increasing the model generalizability. In many cases, the optimal results are obtained using batch sizes smaller or equal to 32 (Masters & Luschi, 2018). Our experiments used batch sizes 32 and 8 consecutively to train the emotion and mental illness detection models. Both models were trained for 15 epochs and used early stopping when the validation loss had stopped reducing. The Adam optimizer (Kingma & Ba, 2015) was used when training both the models with the default learning rate of 0.001.

4.3 Results

4.3.1 Emotion Classification

The emotion detection task was implemented as a multiclass, multi-label classification because the same Twitter message can belong to multiple classes. Since we would like to have independent probability values for each class rather than a probability distribution over the four classes, we used binary cross-entropy as the loss function. Having independent probability values is better oriented towards the identification of independent emotion categories. Table 4.1 reports the emotion classification results obtained using multi-channel CNN (MultiCCNN), CNN with max pooling (CNNMax) and bidirectional Long Short Term Memory module (biLSTM) for comparison.

	Accuracy(%)	**F1(%)**	**P(%)**	**R(%)**	**P ranking(%)**
MultiCCNN	88.88	**77.41**	79.68	75.67	84.85
CNNMax	85.82	68.95	76.97	62.70	81.39
biLSTM	85.07	68.66	75.97	63.49	80.97

Table 4.1: Multi-class, multi-label emotion classification results

In Table 4.1, the recorded accuracy is based on the Keras API accuracy calculation on the multiclass, multi-label models where it takes into account the individual label predictions rather than a full match (i.e. if there is more than one label per instance). The F1 score, precision (P), and recall (R) measures are calculated based on the 'macro' averaging on the exact match, hence the low percentages. We have also reported the label ranking average precision score (P ranking), which averages over the individual ground truth label per instance. This metric is mainly used in tasks that involve multi-label ranking. The results show that the multi-channel CNN model obtained better results than the standard CNN model and the RNN model. Based on this outcome, we have used the above trained multi-channel CNN model to make

predictions on the CLPSych 2015 individual tweets.

4.3.2 Mental Illness Detection

We used binary cross-entropy loss as the loss function and sigmoid function as the last-layer activation function when detecting mental illnesses. The data was sampled using Stratified Shuffle Split to maintain class distribution between 80% of the train and 20% of validation data. Our models were evaluated only on the validation data because the CLPSych 2015 shared task test data labels were not made available. To ascertain the model reliability, we trained the model on five stratified shuffle splits and reported the average of the measured metrics with standard deviation. Considering the results, we identified that having multi-inputs on a multi-task, multi-channel architecture does increase the model's performance.

	Accuracy(%)			**F1(%)**			**Precision(%)**			**Recall(%)**			**AUC(%)**		
	C	D	P	C	D	P	C	D	P	C	D	P	C	D	P
MtMcMi	89.08	87.59	91.35	89.07	83.81	86.06	89.19	86.61	89.81	89.07	82.05	83.50	**95.30**	**92.24**	**93.18**
MtMc	88.55	86.89	91.96	88.53	83.06	87.02	88.82	85.11	90.54	88.54	81.75	84.64	94.62	90.74	92.54
MtMcMiFT	85.50	86.28	91.26	85.45	82.73	85.90	85.91	84.06	89.70	85.49	81.97	83.30	93.88	91.01	91.91
MtMcMiFR	87.42	86.72	91.52	87.41	83.07	86.52	87.48	84.63	89.49	87.41	82.00	84.36	94.88	91.55	92.53
McMclass	92.00	75.69	76.73	89.05	76.83	81.22	86.34	78.25	86.93	92.00	75.69	76.73	92.44	84.55	86.32
biLSTMMtMi	51.35	71.61	78.60	41.27	41.73	44.00	41.92	35.80	39.30	51.52	50.00	50.00	56.87	59.89	60.28
biLSTMMt	52.48	72.31	78.60	47.48	46.40	44.00	47.84	59.81	39.30	52.57	52.06	50.00	56.32	59.96	54.89
svmMclass	81.73	52.91	42.85	75.82	57.01	46.87	70.76	62.11	51.97	81.73	52.91	42.85	81.18	79.12	77.70

Table 4.2: Mental illness detection using multi-task, multi-channel, multi-input architecture

Table 4.2 demonstrates the model performances according to different combinations of the multi-task, multi-channel, and multi-input architectures. To demonstrate the effectiveness of the proposed approach, we conducted several experiments using variants of two deep learning architectures, Convolutional Neural Networks and Recurrent Neural Networks, and to measure a baseline; we used the classical machine learning

approach: Support Vector Machine. The experiments: MtMcMi (Multi-task Multi-channel, Multi-input), MtMc (Multi-task, Multi-channel), MtMcMiFT (Multi-task, Multi-channel, Multi-input, using FastText word representations), MtMcMiFr (Multi-task, Multi-channel, Multi-input, using word Frequencies), McMclass (Multi-channel, Multi-class), biLSTMMtMi (bidirectional Long Short Term Memory, Multi-task, Multi-input), biLSTMMt (bidirectional Long Short Term Memory, Multi-task), svmMclass (support vector machine, Multi-class) were conducted to identify a suitable approach to discover individuals who are susceptible of mental disorders.

The metrics used for the evaluation are accuracy, precision, recall, F1 score, and Area Under the Receiver Operating Characteristic curve (AUC). The reported metrics are from the prediction of users belonging to the three categories, control (C), depressed (D) and PTSD (P). The AUC score is used to compare the model performances, where the average AUC is calculated with standard deviation. The standard deviation is used as a mechanism to identify variance among model performances, which could bring insight into the reliability of the trained model. For each experiment, we recognized that the standard deviation is approximately around 0.01, which provides an empirical confirmation that data sampling using stratified shuffle splits provides an accurate representation of the complete dataset.

4.4 Discussion

The "MtMcMi" architecture uses features based on emotion as multi-inputs on the control and depressed users, while age and gender were used on those with PTSD. In comparison to "MtMc" which is multi-task, multi-channel without multi-inputs, we could see that using multiple inputs has increased the average AUC score and most

of the other evaluation metrics (i.e., precision, recall, and F1 score). Even though the increase could not be considered significant, the potential room for improvement is high if provided with more accurate emotion prediction and additional profound features identified by researchers. Concerning the emotion detection task, we could have improved prediction accuracy if provided with additional data when training the deep learning model. It is essential to highlight that our research objective is to investigate the opportunities that can be derived by combining features generated using automatic and manual methods.

When analyzing the result for "MtMcMiFr", which uses the Multi-task, Multi-channel, Multi-input architecture with the vocabulary created using the default word frequencies, we could identify our proposed approach, which uses the vocabulary constructed using the weighted TF-IDF words has produced comparatively better results. Even though the gained improvements could not be considered significant, the proposed method can be an effective approach when initiating the vocabulary that balances the rare and frequently used words.

In the experiment "MtMcMiFT", where we used the fastText embeddings layer with the number of dimensions equal to 100 as an input to the Multi-channel convolutional neural network, it was observed that the results obtained using the randomly-initialized embedding layer are better than the results with the fastText pre-trained embeddings. As mentioned in the background section, this could be due to various reasons such as embeddings not being trained on a sufficiently large dataset where we trained it only on the CLPSych 2015 dataset or the negative impact of using the most similar word for missing words. In future work, we will conduct further research to enhance the embedding layer word representation by training the fastText model using

additional data and a combination of different hyperparameters. The effectiveness of using convolutional neural network models can be identified when evaluating the results obtained using Recurrent Neural Network (RNN) based architectures. The "biLSTMMtMi" method uses a bidirectional Long Short Term Memory ("biLSTM") model in a Multi-task, Multi-input design and comparatively has produced poor results. This could be due to several reasons, such as the unstructured nature of the Twitter text and the non-exitance of long-term dependencies. For example, our best results were obtained when using the kernel sizes (i.e., the number of consecutive tokens) one, two, and three, and once the kernel sizes are increased, the overall model predictability decreases. When using a "biLSTM" model as the shared layer in multi-task learning without multiple inputs ("biLSTMMt"), the results are somewhat better compared to when using multi-inputs.

To demonstrate the effectiveness of multi-task learning to detect multiple mental disorders, we compared the proposed approach with multiclass classification to distinguish neurotypical users from users susceptible to PTSD or depression. We used a multi-channel convolutional neural network to predict the three classes (i.e., control, depression, and PTSD). Compared to our proposed approach, we can identify that multiclass classification using CNN has produced slightly better results on two occasions, which is for average accuracy and recall under the control class. Through further analysis, we see that average precision, F1 score, and AUC scores are higher for all three classes when using the proposed approach. Overall the multiclass classification task has produced low scores (especially for precision, F1 score, and AUC) when detecting users susceptible to depression and PTSD, while the proposed approach has contributed significantly better results. The better results could be

because depression is commonly identified among individuals with PTSD, and the shared layer has managed to learn such common characteristics while the task-specific layers have learned the individual features unique to each disorder.

We used the linear SVM classifier with TF-IDF features (200,000) in a multiclass classification task as a baseline. When sampling the data, five splits of 80% training and 20% testing were created using the Stratified Shuffle Split method to maintain class distribution.

4.5 Comparison to Related Work

Even though our results cannot be compared directly with the CLPSych 2015 task participant results, we can identify that our proposed model has produced competitive results when comparing the AUC scores for detecting users with PTSD and depression. Using our proposed architecture, we obtained an AUC score > 0.90 in identifying users belonging to the control, PTSD and depression categories. The CLPSych 2015 shared task best results obtained by Resnik, Armstrong, Claudino, and Nguyen (2015) have reported AUC scores of 0.86 (depression vs. control), 0.84 (depression vs. PTSD) and 0.89 (PTSD vs. control) and similarly, Preotiuc-Pietro et al. (2015) have reported an average AUC of 0.86 in differentiating neurotypical users from users susceptible to PTSD and depression. The best results reported by the task participants are on the test dataset, which was made available only to the registered participants and not to the public. Due to this reason, our predictions were made only on the validation data, which was averaged over five stratified shuffle splits with a standard deviation around 0.01. Benton, Mitchell, and Hovy (2017) used multi-task learning to predict mental disorders and suicide ideation and reported an AUC score < 0.80 for predicting

users diagnosed with PTSD or depression and an AUC score > 0.90 for detecting neurotypical users. As we have also used multi-task learning within our proposed solution, our model's prediction results can be considered competitive. Similar to when comparing with the best results of CLPSych 2015, we could not directly compare our results with Benton, Mitchell, and Hovy (2017), due to the combination of datasets used by the authors.

4.6 Summary

Overall, we can see that using limited unstructured data with an architecture based on CNN has produced better results compared to RNN-based architectures. Notably, the multi-task, multi-channel architecture with multiple inputs has provided the best results and confirms that using multiple inputs positively influences the overall model performance. Also, the appropriateness of using multi-task learning instead of multi-class classification to detect multiple mental disorders is highlighted. Similar to the fact that certain mental disorders share specific common symptoms (American Psychiatric Association, 2013), multi-task learning has managed to learn such characteristics through a shared representation followed with task-specific layers to identify the unique attributes to differentiate between multiple mental disorders.

Chapter 5

Suicide Ideation and Mental Illness Detection using Multi-Task Learning

With the success of detecting multiple mental disorders using multi-task learning, we investigate the possibilities of using multi-task learning to simultaneously predict users with suicide ideation and mental disorders. In recent years, there has been an increase in research that uses deep learning methods to detect suicide ideation and mental disorders using publicly available social media data. However, to the best of our knowledge, not much research has used multi-task learning models to detect mental illness or suicide ideation. Also, we have not identified any research on detecting suicide ideation and mental disorders using multi-task learning with two different datasets submitted in parallel to a single model. Given the challenges faced when creating a dataset to predict mental disorders (Coppersmith et al., 2014b) or suicide ideation (Shing et al., 2018), it will be complicated to construct a dataset with the same users with mental disorders and suicide ideation. Also, having fewer datasets with limited data points makes using multiple datasets in parallel with MTL a more viable solution.

However, not much research has used multi-task learning models to detect mental illness or suicide ideation. Also, we have not identified any research on detecting suicide ideation and mental disorders using multi-task learning with two different datasets submitted in parallel to a single model. Given the challenges faced when creating a dataset to predict mental disorders or suicide ideation, it will be complicated to define a dataset that captures a single user with mental disorders and suicide ideation.

This chapter will address several vital objectives to validate the proposed architecture. The key objective of the research is to predict users with suicide ideation and a mental disorder by feeding two different datasets in parallel into a multi-task learning architecture. In order to identify the best approach to train a multi-task learning model given the two datasets, we performed several experiments using different multi-task learning architectures. After selecting the most appropriate MTL architecture, we performed suicide ideation and mental illness detection using several combinations of datasets. The combinations are based on the self-reported mental disorders where certain users are diagnosed with either a single or multiple mental illnesses (also known as a comorbidity of mental disorders). To understand the level of impact mental disorders have on suicide ideation, we divided the experiments into two streams where one is to predict users with suicide ideation or a mental disorder using the posts from users with a single mental disorder, and the second stream is to use the posts from users with multiple mental disorders.

To validate the effectiveness of the proposed architecture, we conducted the experiments in line with the CLPSych 2019 shared task, where the objective of the shared task is to predict the level of suicide risk users has revealed through their posts published in the Reddit social media platform. The main objective of the current

research is to identify the impact mental disorders have on suicide ideation; hence, we have focused mainly on two subtasks from the CLPSych 2019 shared task. Rather than detecting the level of suicide risk, we will focus on identifying whether a user has a suicide risk, that is, the task of distinguishing users with "no risk" from the users with "low", "moderate", and "severe" risks combined. The second subtask that our experiments are going to be based on is detecting users with suicide ideation that requires urgent attention from the users who do not, and that is to distinguish users with "no" or "low" risks from the ones with "moderate" or "severe" risks.

5.1 Data

5.1.1 UMD Dataset

We used the dataset provided by the CLPSych 2019 shared task. Table 3.6 demonstrates in-depth details about the dataset being used with the associated class distributions. For the binary classification task of predicting "flagged/not flagged" users, 127 users with "no" risk are combined with 242 randomly selected users from the control group to form the negative class. The total number of users in the positive class that is by combining the classes: "low", "moderate", and "severe" are 369 users. For the binary classification task in detecting "urgent/not urgent" users, 177 users from the combined classes, "no" and "low" risks were merged with 142 randomly selected users from the control group to form the negative class. For the positive class, a total of 319 users were selected from the "moderate" and "severe" risk groups. The test dataset provided by the CLPSych 2019 task organizers is taken without any modifications to evaluate our proposed architecture against the state-of-the-art results produced by the task participants. The test dataset constitutes 125 users in order of 32, 13,

28 and 52 from the classes “no”, “low”, “moderate”, and “severe”. After combining the classes according to the two tasks, the test dataset for the “flagged/not flagged” task contained 93 users in the positive class and 32 in the control group. For the “urgent/not urgent” task, 80 users were included in the positive class, while 45 users were merged from the “no” and “low” risk groups to form the control group.

Even though the task organizers have released the train and test datasets with equal class weights, we could not maintain the same distributions due to architectural requirements. Therefore, even though the training dataset that we created did not have the same class distribution as the test dataset, comparing our test results against the results of the task participants, we can identify that our trained models are well generalized with scope for further improvements.

5.1.2 SMHD Dataset

According to table 3.8, the data mentioned in the single and multiple disorder columns were both used for the flagged/not flagged and urgent/not urgent tasks. In the flagged/not flagged task, we used the data from the single disorder column to identify the impact on suicide ideation detection from the users who self-declared a single mental disorder. In converse, the data from the multiple disorders column are used to ascertain the shared feature space between the users with suicide ideation and multiple mental disorders. Because one of our key research objectives is to identify the impact mental disorders have on suicide ideation detection, we have centred the sample selection around the dataset dimensions made available by the CLPSych 2019 shared task. Due to the inadequate number of samples in the SMHD dataset, that is, not having an adequate number of instances to match with the suicide ideation

detection task, we did not use the users who self-reported "eating disorders". There must be at least 399 users (319 users for training and 80 for testing) to fulfil the binary classification task of urgent/not urgent suicide ideation detection.

To predict users into either flagged or not flagged classes, first, we used users who have self-declared a single mental disorder and randomly selected a sample to match the total number of users having suicide ideation (i.e., 369 users). From the remaining users, we randomly selected 93 users for testing. Similarly, we randomly selected 369 control users from the related random control group (i.e., a combined control group from the train, validation and test) for training and from the remaining control users, another 32 users were selected for testing purposes. For example, when feeding the model with users who self-declared bipolar disorder, we will first select a random sample of 369 out of 2,094 users and from the remaining 1,725, a random sample of 93 users will be selected for testing. The same approach is taken for the users with multiple mental disorders except that we filtered the users who have self-declared bipolar disorder and one or more of the remaining eight mental disorders. According to table 3.8, there are 1,546 users with multiple mental disorders, including bipolar disorder, and 369 users will be randomly selected for the positive class. From the remaining 1,177 users, another 93 users will be randomly selected for testing purposes as the positive class. Equally, the same number of users selected for training and testing will be selected for the control groups.

The same approach as above will be followed when predicting users with suicide ideation and requiring urgent attention. The only difference will be the number of users where for training and validation purposes, 319 users will be filtered, and from the remaining users, another 80 will be randomly selected for testing. Thus, the exact

number of users as the positive class (i.e., 319 users) will be selected for the control group, while 45 users will be selected for testing purposes. When there are not enough samples in either single disorder or multiple disorder categories, the samples were combined to form both categories to construct the required dataset. For example, when experimenting with users who have self-declared schizophrenia and one or more additional mental disorders, the number of users available is not enough to be used with the proposed architecture. The urgent/not urgent binary classification task must have 319 users for the positive class where only 299 users who self-declared schizophrenia was available. To overcome such situations and to submit the dataset according to the model requirements, we randomly selected the remaining 20 users from the category of users with a single disorder. Similarly, users with Autism did not have an adequate number of test users with multiple mental disorders, where for the tasks flagged/not flagged, the remaining users were randomly selected from the single mental disorder group.

5.2 Model Architecture

5.2.1 Proposed Architecture

Continuing from multi-task learning to detect multiple mental disorders, we investigated the feasibility of implementing a similar architecture to detect mental disorders and suicide ideation. In chapter 4, we used a single dataset to detect multiple mental disorders where a user was labelled with one condition. For example, the CLPSych 2015 Twitter dataset contained users with either PTSD or Depression, and none of the users were annotated for having both the disorders. The ones without any of the disorders were categorized into the control group. The main objective of our

experiments is to identify whether certain mental disorders and suicide ideation share common features and, if so, whether we can use the shared feature space to enhance the model performances. Unlike related literature where datasets were combined to form a single dataset with users having mental disorders or have attempted suicide (Benton, Mitchell, & Hovy, 2017), we did not follow the same approach but submitted the datasets independently to our proposed model to predict users with mental disorders or suicide ideation. Hypothetically, combining users with mental disorders and suicide ideation (i.e., if the initial annotations were only for a single task and the combined datasets were not re-annotated accordingly) could introduce invalid narratives where it has been identified that users with suicide ideation could be diagnosed with different mental disorders such as PTSD (LeBouthillier et al., 2015; Wilcox et al., 2009), bipolar disorder (Dome et al., 2019; Simpson & Jamison, 1999) and mood disorders such as depression (Bertolote & Fleischmann, 2002; Bertolote et al., 2004). Also, when combining datasets containing users with mental disorders and suicide ideation, certain users within the suicide dataset could have been diagnosed with either one or more mental disorders. Such circumstances could allow the model to penalize features strongly associated with suicide ideation and mental disorders based on the prediction outcome.

We conducted several experiments using different multi-task learning architectures such as hard parameter sharing (figure 2.3), soft parameter sharing (figure 2.4), and mixed parameter sharing (figure 2.5). However, the model trained using the mixed parameter sharing architecture (i.e., soft and hard parameter sharing) produced better results on unseen data. The architecture of the best-performing model is mentioned in figure 5.1.

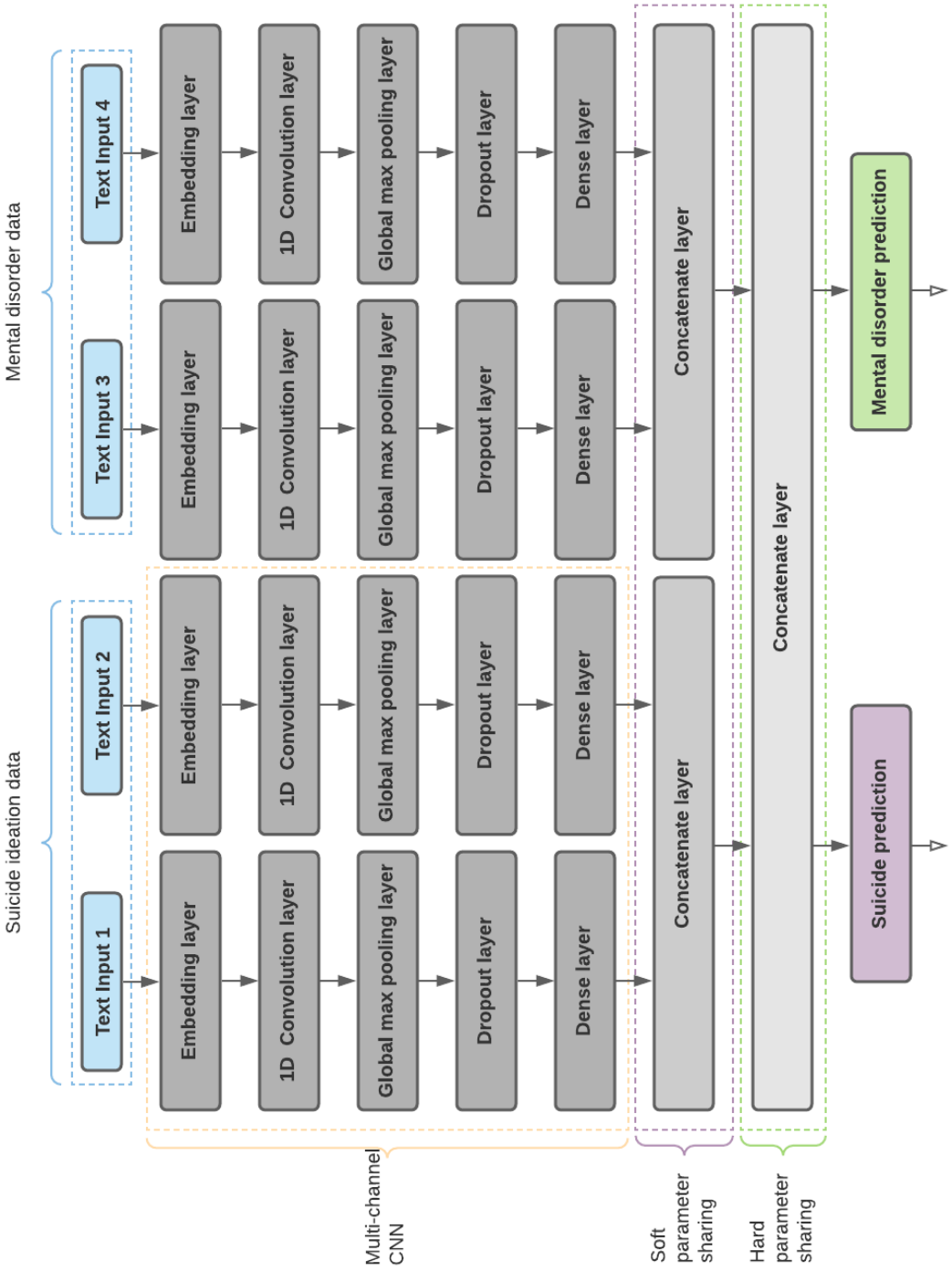


Figure 5.1: Proposed multi-task learning architecture with mixed parameter sharing

We used a multi-channel Convolutional Neural Network (Y. Kim, 2014) as the core computational unit in our proposed architecture. However, unlike in the architecture proposed for detecting multiple mental disorders using CLPSych 2015 Twitter data, we used only two channels to reduce the model complexity and lessen the negative impact of model overfitting given the limited training data. We also discovered that using a CNN architecture with a single channel underfit the training data and performed poorly on the holdout datasets.

For each CNN channel processing suicide ideation and mental disorder data, we used two different kernel sizes so that input posts will get processed with different n-gram (i.e., sequence of words) sizes. Each CNN channel was initialized with kernel sizes 3 and 4 (i.e., for each CNN layer taking the input from either suicide data or mental illness data). In search of optimal kernel size, we experimented with several values, in which the kernel sizes 3 and 4 produced better results than sizes 1 and 2, 2 and 3, and 4 and 5. Each of the one-dimensional convolution layers contained 256 filters with the selected kernel size.

To prepare the input for the CNN layer, we used a randomly initialized and trainable embedding layer with an output dimension of 300 units. Using different dimensions for the dense embeddings such as 100, 200 and 300, we identified 300 units as the optimal embedding dimension that could be used to enhance model predictability. Each of the convolution layers was further parameterized with a stride (i.e., the distance between the successive sliding windows) of 1 and "valid" padding instead of "same" padding. We used valid padding so that no padding is applied to the input as opposed to the "same" padding where input and output dimensions are kept the same. In addition, we conducted several experiments with

different pre-trained word embedding architectures such as fastText (Grave et al., 2018), GloVe (Pennington et al., 2014), Byte-Pair embeddings (Heinzerling & Strube, 2018), Character Embeddings (Lample et al., 2016) and stacked embeddings (Akbik et al., 2018). We determined that randomly-initialized embeddings produced better results for our tasks, than the pre-trained embedding architectures even when the pre-trained weights were fine-tuned during model training. However, when comparing the model performances using different pre-trained embeddings, the model trained using fastText embeddings generalized well on the unseen data. The reason that randomly initialized embeddings produced better results could be because most of the pre-trained embeddings were trained on the text that is not closely related to the text used in suicide ideation and mental illness related posts published in the Reddit social media platform. We did not train any embedding architecture with our data as our main objective is to identify the impact different mental disorders have on suicide ideation.

The input dimension or the vocabulary size differed based on the input text, where the vocabulary was created using the training data from both SMHD and UMD datasets. For example, when the two tasks in the multi-task learning model predict users with and without suicide ideation and users with and without PTSD, the vocabulary is created by joining the training data from the UMD and SMHD datasets (i.e., filtered data with users having PTSD). Creating the vocabulary from both datasets allows the tasks to converge into a shared feature space that could improve the model performances.

The output from the convolution layer is sent through a Global Maximum Pooling layer to reduce the number of learnable parameters, and as a result, it could reduce

model overfitting. Furthermore, by reducing the number of learnable parameters, the computational overload can also be reduced.

We took several measures to regularize the network so that the impact of model overfitting when using a limited number of data points can be reduced. We used dropout (Srivastava et al., 2014), where the output from the CNN layers will be randomly assigned with zeros and L1 and L2 regularization to penalize layers for having larger weights. The regularized output from the CNN layer was sent through a fully connected layer with 512 hidden units. The dense representation from each channel is (i.e., two channels for each task) merged to form a task-specific representational vector. The merged vectors are used in the process of soft parameters sharing. The distance between the parameters of the two tasks was regularized to identify the similarity between the tasks. For example, to identify the similarity between the two tasks, the task parameters are regularized using mean squared error.

$$MSE = \frac{1}{N}\sum_{j=1}^{N}(V_{T1}^{(j)} - V_{T2}^{(j)})^2 \tag{5.1}$$

V_{T1} = parameter vector of task 1

V_{T2} = parameter vector of task 2

j = j'th parameter

N = number of parameters (in the proposed model it will be 1,024)

Each of the task-specific representational vectors is concatenated to form the shared representation layer containing 2,048 features. The shared representation is used as an input to two softmax layers to generate the class probabilities for each task. Hypothetically sharing parameters between the tasks could reduce overfitting where

rather than fine-tuning parameters on a single task, the parameters are optimized on multiple tasks (Ruder, 2017).

5.2.2 Proposed Architecture with Auxiliary Inputs

To identify whether or not auxiliary inputs can be used to enhance the model performances, we extended the proposed architecture mentioned in figure 5.1 to be tested with multiple inputs discovered during the exploratory analysis stage (see section 3.3.1).

In addition to the text inputs from the users with suicide ideation and mental disorders, the EMPATH categories extracted from the content posted by users related to suicide ideation and mental illnesses are used as an auxiliary input to identify the impact of such features in predicting users with suicide ideation and mental disorders. After scaling the extracted EMPATH features, we transformed the given auxiliary input by sending it through a densely connected layer. The number of hidden units used with the dense layer differed based on the task. When predicting users with and without suicide ideation and mental disorder, we used 8 hidden units, and if applied with more units, the model performances decreased. To detect users with suicide ideation who require urgent attention, we identified 16 units as the optimal number of hidden units. However, when conducting the experiments to prove the domain adaptation capabilities of the proposed model where Reddit posts were used to detect suicide ideation and tweets as the input to detect users with mental disorders, we used 32 hidden units with the dense layer to transform the auxiliary input. The transformed data is merged with the multi-channel CNN outputs generated using suicide ideation and mental disorder data. Similar to the proposed architecture, the soft parameter

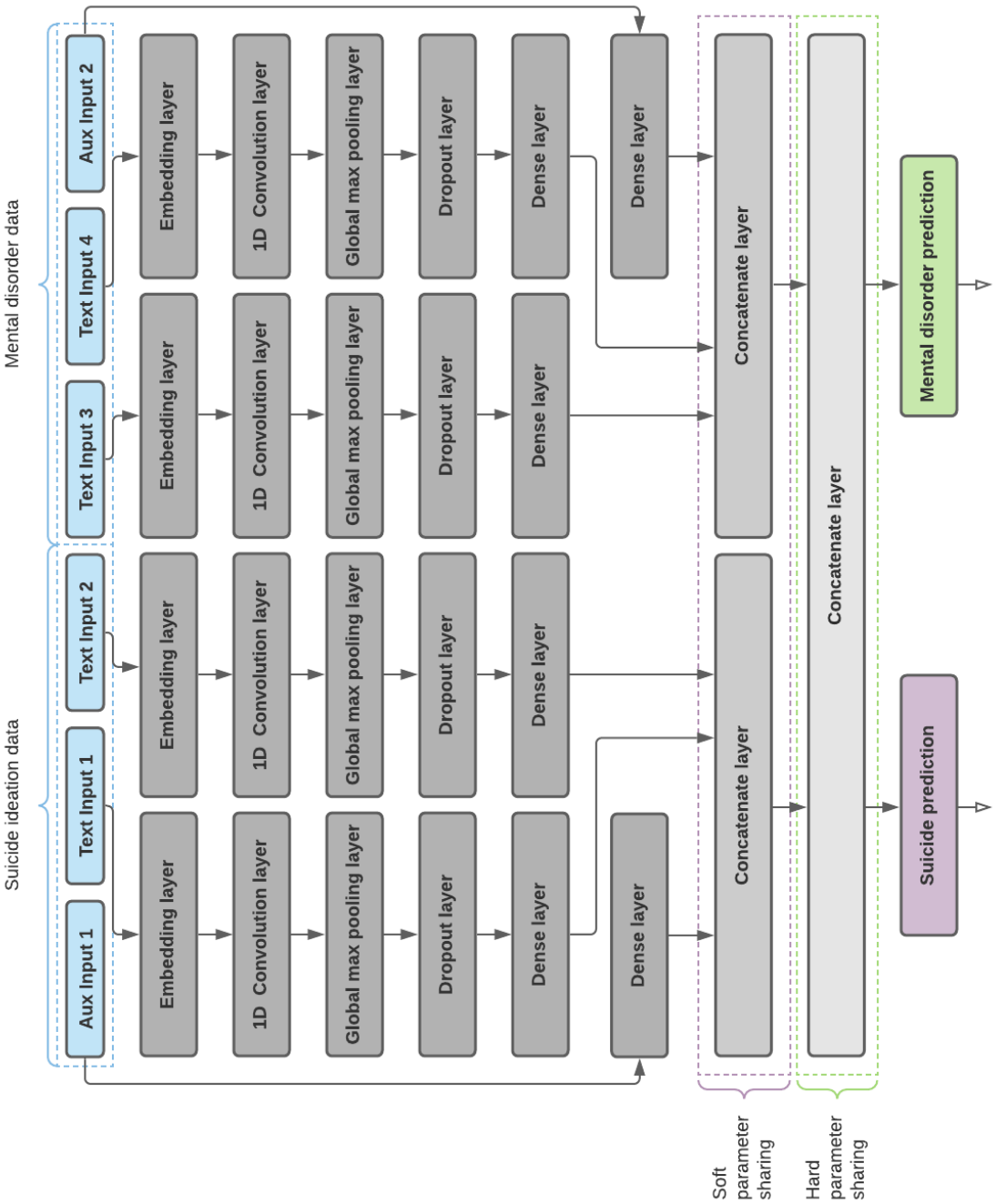


Figure 5.2: Proposed multi-task learning architecture with mixed parameter sharing and auxiliary inputs

sharing will be based on the vectors created using the transformed inputs using the suicide ideation or mental illness data concatenated with the transformed auxiliary inputs. Given the task of detecting users with suicide ideation and mental disorders, the parameter vectors to be regularized will have 1,032 parameters. The vectors merged with the auxiliary inputs will be concatenated to form the hard parameter sharing layer.

5.2.3 Baseline Architecture

We used a multi-channel CNN with a similar configuration as in the proposed architecture but without multi-task learning to use it as a baseline for comparison (single task learning).

A similar number of parameters to the proposed architecture is used for the baseline architecture, which contained a randomly initialized embedding vector of 300 dimensions connected to a one-dimensional convolution layer with 256 filters. The multi-channel CNN network used the kernel sizes 3 and 4 with global maximum pooling and dropout for network regularization. The outputs from each channel are concatenated and submitted to a softmax layer to generate the class probabilities.

5.3 Experiments

The main objective of our research is to ascertain the impact mental disorders have on suicide ideation by conducting experiments to identify if users with suicide ideation share a hidden feature space with users diagnosed with one or more mental disorders. During our research, we investigate the impact different mental illnesses have on suicide ideation and whether users diagnosed with multiple mental disorders share

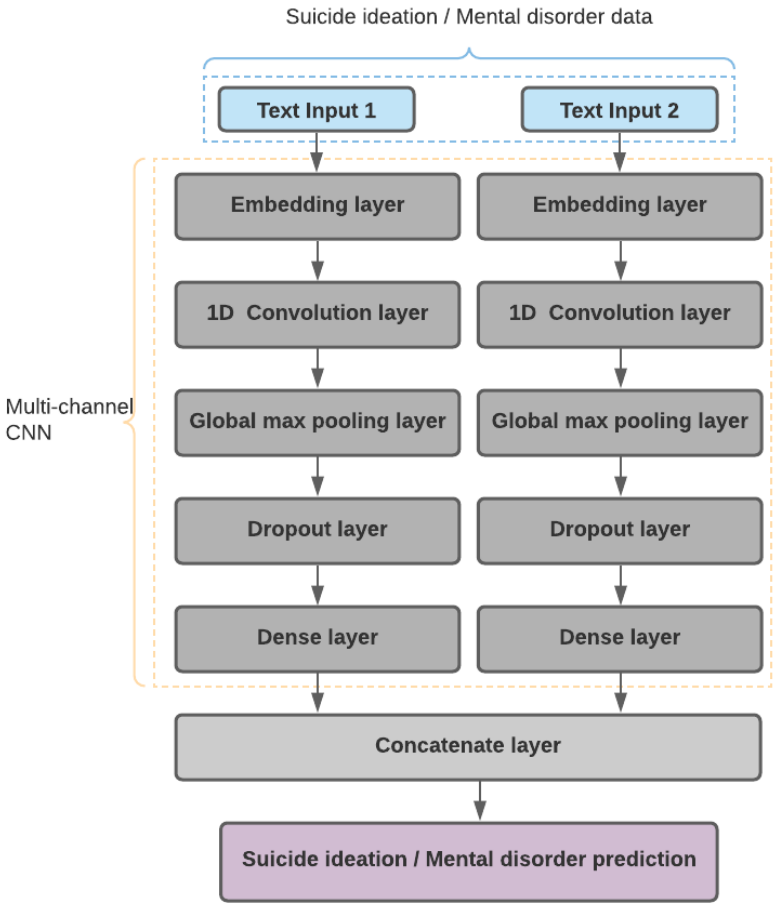


Figure 5.3: Baseline architecture for single-task learning

more features with suicide ideation than those diagnosed with a single mental disorder. In addition, we also investigate the impact suicide ideation detection has on mental illness detection to identify the disorders that share more hidden features with suicidal thoughts. All the data used in predicting suicide ideation is the same as the CLPSych 2019 shared task data, and to demonstrate the generalizability of our proposed model further, we used the data annotated by the experts, which is also a part of the UMD dataset but not made available to the CLPSych 2019 shared task participants. In addition, we conduct several experiments to demonstrate the impact auxiliary inputs

have on suicide ideation and mental disorder detection. The auxiliary inputs were identified through exploratory analysis and will be further analyzed in future work to understand the extensive impact these features have on suicide ideation and mental disorder detection. We used similar experiment structures for both the research streams flagged/not flagged and urgent/not urgent where one is to detect users with and without suicide ideation (flagged/not flagged), and the second is to identify users with suicide ideation who require urgent attention (urgent/not urgent).

5.3.1 Task: Flagged/not Flagged

As mentioned in subsection 3.1.3 and also in table 3.6, we merged users in different risk categories according to the given task to generate the positive and the control user groups. Each class consisted of 369 users. As mentioned in subsection 3.1.4, we randomly sampled 369 users from each mental disorder group (i.e., from the eight mental disorders) to create the positive class for the mental illness detection task. As for the control group, a matching number of users from the provided control dataset were randomly selected. Each task in the multi-task environment will be a binary classifier, and a total of eight classifiers will be created for each primary diagnosis. To evaluate the trained model, we use the test data made available for each task. For suicide ideation detection, we use 125 users (i.e., 32 users for the control class and 93 users for the positive class) provided by the CLPSych 2019 shared task organizers. We select a random test set for mental illness detection equal to the number of samples made available for suicide ideation detection (i.e., 125 users).

To measure the impact word embeddings have on model performances, we used pre-trained fastText word embeddings with 300 dimensions. We made the embedding

weights trainable so that the parameters of the embedding layer could get adjusted according to the task. We conducted several experiments to identify the impact EMPATH categories discovered using the exploratory analysis have on model performances when used as an auxiliary input. We identified that the top 14 EMPATH categories (i.e., based on the most frequently used terms) discovered during exploratory analysis enhanced the model performances. The categories used as auxiliary inputs are as follows:

'health', 'medical_emergency', 'sadness', 'nervousness', 'fear', 'contentment', 'domestic_work', 'horror', 'torment', 'suicide_topw', 'neglect', 'shame', 'suffering', 'sexual'

It is important to note that we did not conduct extensive experiments using different combinations of EMPATH categories. Our main objective is to identify the impact mental disorders have on suicide ideation detection by predicting users with suicidal thoughts and mental disorders in a multi-task learning environment. Also, we tested the auxiliary inputs only on the best-performing model.

5.3.2 Task: Urgent/not Urgent

When selecting the data to predict users with suicide ideation that requires urgent attention, the same procedure as before when preparing the data for the flagged/not flagged task is followed. The only difference between the two tasks is the sample size where users from both "No" risk and "Low" risk categories are merged into creating the control group. According to Tables 3.4 and 3.6, we selected 319 users from both "Moderate" and "Severe" risk categories to create the positive class. The negative class contains users from the "No" risk and "Low" risk categories that add up to 177

users. We randomly selected 142 users from the UMD control group to balance the negative class with the positive class. Similar to the flagged/not flagged task, we selected two sets of user groups with single or multiple mental disorders. Each user group contained 319 users for the positive class and 319 users for the control group. The test dataset for suicide ideation detection was kept untouched, where it contained 125 users, 45 users in the control group and 80 users labelled as positive. The same number of data points as the UMD test dataset is randomly selected from the SMHD dataset during inference (i.e., using unseen test data on the trained model).

Like the flagged/not flagged task, we conducted several experiments using the pre-trained fastText word embeddings with 300 dimensions. To identify the impact of EMPATH categories on predicting users with suicide ideation or a mental disorder, we experimented with several categories picked from the top 14 EMPATH categories (i.e., filtered based on term frequency) discovered during exploratory analysis.
'neglect', 'anger', 'sadness', 'torment', 'emotional', 'shame'
Even though we could not identify any improvement over the results when not using the EMPATH categories, more opportunities are there to conduct extensive experiments using different combinations of the categories to enhance model performances. Therefore, for our future work, we will conduct further experiments to identify combinations of EMPATH categories that can improve the model performances.

5.3.3 Creating the Vocabulary

To create the vocabulary, we merged the UMD and SMHD training data. Taking the most frequent tokens from both the training datasets allows the model to be generalized with a fair representation of content from users having suicide ideation and

mental disorders. The merged dataset was used to compute the maximum sequence length so that longer sequences could be normalized. We calculated the mean sequence length from the merged training datasets and selected the maximum sequence length, five standard deviations away from the average length. The reason for taking five standard deviations is to cover as many users as possible without truncating any text from the users' concatenated posts. The mean sequence length will change between each classification task, and as a result, the maximum sequence length will also change. For example, when detecting users with suicide ideation and PTSD (i.e., with a single mental disorder), the mean sequence length was 3,503 tokens with a maximum sequence length of 49,964 tokens. A total of 518 users recorded a sequence length longer than the mean sequence length. To calculate the new maximum sequence length, we took five standard deviations away from the mean and identified the new sequence length to be 24,553 tokens, which covered a total of 99.39% of users from the merged training dataset. We identified nine users with a longer sequence length than the new sequence length from the training dataset. The nine users comprised of one user with PTSD, one user from the mental disorder control group, two users from the suicide ideation control group and five users from the suicide ideation positive class. To ascertain a better approach to normalize the sequence lengths that are longer than the calculated maximum sequence length, we completed several experiments by removing tokens from the "front", "end", and "randomly" from the given sequence of tokens. Our experiments identified that better performances were obtained for the flagged/not flagged task when normalizing the longer sequences by removing tokens from the "end" of the sequence, while for the task of urgent/not urgent, it is from the "front". Due to the reason that only a few users were identified for having longer sequences than the

newly calculated maximum sequence length, it is difficult to establish an argument to justify that the most prominent features that distinguished users with suicide ideation or mental disorders were identified not from the end of the tokenized sequence. We took all the text without truncating the sequences to generate the vocabulary with unique tokens. We did not use the same method that was used to generate the vocabulary when detecting users with multiple mental disorders (see section 4.2.1) due to the reason that our main objective is to identify the level of impact mental disorders have on suicide ideation detection without using other external factors that could enhance the model performances. The size of the vocabulary changed with each classification task and for each research question. For example, all the tokens from both training datasets were used to generate the vocabulary when detecting suicide ideation or the mental disorder PTSD (i.e., for a single mental disorder). As a result, the total number of tokens used in the vocabulary was 119,691. Tables 5.1 and 5.2 demonstrate the maximum sequence length, average sequence length, new maximum sequence length, number of users having a sequence length below the new maximum sequence length and the vocabulary size.

According to Tables 5.1 and 5.2, the maximum sequence length is 49,964 identified from the suicide ideation dataset. On average, the new maximum sequence length for the task flagged/not flagged is around 24,182, with an average vocabulary size of 124,222. The urgent/not urgent task has an average new maximum sequence length of 24,928 with an average vocabulary size of 114,609. Therefore, taking five standard deviations away from the mean has resulted in an effective maximum sequence length rather than the maximum sequence length identified from the suicide ideation detection data that could substantially increase model training time and memory requirements.

Tasks	Maximum sequence length	Average sequence length	New maximum sequence length	Users under the new maximum length(%)	Vocabulary size
Suicide + PTSD: single / multiple	49,964	3,503.60	24,553	99.39	119,692
	49,964	3,524.53	25,185	99.39	124,679
Suicide + Anxiety: single / multiple	49,964	3,350.81	23,699	99.45	123,625
	49,964	3,427.14	24,929	99.39	126,007
Suicide + Depression: single / multiple	49,964	3,325.31	23,914	99.45	125,141
	49,964	3,293.76	23,724	99.39	124,759
Suicide + Bipolar: single / multiple	49,964	3,352.45	24,091	99.32	121,426
	49,964	3,404.37	24,181	99.39	123,278
Suicide + OCD: single / multiple	49,964	3,377.05	24,392	99.52	123,375
	49,964	3,401.73	24,354	99.45	123,545
Suicide + Schizophrenia: single / multiple	49,964	3,253.20	23,916	99.45	125,819
	49,964	3,317.45	23,492	99.45	126,738
Suicide + ADHD: single / multiple	49,964	3,326.01	23,559	99.45	121,135
	49,964	3,523.33	24,500	99.45	125,750
Suicide + Autism: single / multiple	49,964	3,383.49	24,303	99.39	127,096
	49,964	3,394.33	24,117	99.32	125,482

Table 5.1: Merged training data statistics for the task flagged/not flagged

Tasks	Maximum sequence length	Average sequence length	New maximum sequence length	Users under the new maximum length (%)	Vocabulary size
Suicide + PTSD: single / multiple	49,964	3,589.39	25,221	99.29	111,076
	49,964	3,613.77	25,956	99.29	115,066
Suicide + Anxiety: single / multiple	49,964	3,389.28	24,088	99.37	109,781
	49,964	3,519.29	25,872	99.29	117,229
Suicide + Depression: single / multiple	49,964	3,424.37	24,730	99.37	114,729
	49,964	3,403.60	24,534	99.29	116,224
Suicide + Bipolar: single / multiple	49,964	3,477.45	25,097	99.21	112,141
	49,964	3,488.75	25,020	99.29	113,274
Suicide + OCD: single / multiple	49,964	3,480.53	25,290	99.45	113,972
	49,964	3,503.16	25,214	99.37	114,611
Suicide + Schizophrenia: single / multiple	49,964	3,357.24	24,485	99.37	118,173
	49,964	3,334.11	23,713	99.45	114,204
Suicide + ADHD: single / multiple	49,964	3,446.24	24,409	99.37	111,642
	49,964	3,604.46	25,213	99.45	116,449
Suicide + Autism: single / multiple	49,964	3,472.25	25,095	99.29	117,908
	49,964	3,503.65	24,904	99.21	117,266

Table 5.2: Merged training data statistics for the task urgent/not urgent

5.3.4 Baseline

To calculate the baseline, we separated each task of the multi-task learning environment and trained different models using a multi-channel CNN network as mentioned in section 5.2.3 (figure 5.3). For suicide ideation detection, we used the same dataset used with the multi-task learning environment for the two different tasks: flagged/not

flagged and urgent/not urgent. For the flagged/not flagged task, we used 369 users in the positive class, which contains users from the "Low", "Moderate", and "Severe" risk groups and the same number of users for the negative class comprising users from the "No" risk category and randomly selected users from the control group to balance the difference between the positive class and the "No" risk group. Similarly, for the urgent/not urgent task, we selected 319 users for the positive class, a combination of "Moderate" and "Severe" risk categories. The negative class comprised users from the "No" and "Low" risk categories combined with randomly selected control users to gap the difference between the positive and the negative classes. The test data was used without modification (i.e., negative class not being balanced by merging with randomly selected control users). The test data split for the flagged/not flagged was 93 / 32, and for urgent/not urgent, it was 80 / 45.

In addition to using the crowdsourced data during inference, we used the UMD expert annotated data to calculate a baseline to better understand how well our model trained on crowdsourced data has generalized on the expert annotated data.

To calculate a baseline for the users who self-declared single or multiple mental disorders, we experimented: for each task (i.e., for the tasks of flagged/not flagged and urgent/not urgent), for each mental disorder (i.e., for eight mental disorders) and also for multiple or single mental disorder diagnosis. For example, in the flagged/not flagged task, to predict if a user self-declared PTSD (i.e., single mental disorder), we select a random sample of 369 users from a group of users who self-declared only PTSD. To match the positive class, we randomly select 369 users from the control group. At inference, we randomly select 93 users who self-declared only PTSD and 32 users from the control group. Similarly, to calculate the baseline for users with

multiple diagnoses, we select a similar random sample as before but only from a group of users who have self-declared PTSD and one or more other mental disorders. For the urgent/not urgent task, the changes will only be on the sample size where the training dataset will contain 319 users, each in the positive and the negative classes. At inference, the positive class will contain 80 users and 45 users in the control group.

5.3.5 Model Training

The data frame containing the pre-processed and normalized (i.e., having a uniform sequence length) data was split into five stratified shuffle splits (Pedregosa et al., 2011) with 80% of data for training and the remaining 20% for validation. The most critical aspect to consider when submitting the input data to the model is the task alignment given the data splits. For example, when fitting the model with the input data, the positive and the negative classes from both the datasets must be aligned so that a user with suicide ideation is in parallel with a user diagnosed with a single or multiple mental disorder.

Once the input data is submitted accordingly, soft parameter sharing will be on the parameter vectors generated using data points belonging to users with suicide ideation and mental disorders. If the tasks are misaligned, the comparison will be based on users with suicide ideation and users from the mental disorders control group or the other way around. During the hard parameter sharing stage, the two tasks will share from the concatenated feature vectors derived from the users with/without suicide ideation and with/without a single or multiple mental disorders. When training the model, it was identified that the most stable learning rate is 0.001 with Adam optimizer (Kingma & Ba, 2015). The Rectified Linear Unit (ReLU) (Nair & Hinton, 2010) was

applied to the output generated by both the CNN layers and the Dense layers that follow. To reduce model overfitting, we used dropout (Srivastava et al., 2014) with a probability of 0.5 and $L1$ and $L2$ regularization with a regularization factor (10^{-5}) to penalize convolution and fully connected layers for having larger weights. When fitting the data to the proposed architecture, we made sure not to shuffle data to maintain task alignment. We used a custom loss function, which summed categorical cross-entropy loss and mean squared error. The mean squared error was used to regularize the parameter vectors of the two tasks (i.e., suicide ideation detection and mental illness detection). When training both flagged/not flagged and urgent/not urgent tasks, we experimented with several mini-batch sizes and identified that smaller batch sizes produce better results than larger batch sizes (Masters & Luschi, 2018). A batch of sizes: 8, 16 or 32 were used in the experiments and a mini-batch of size 8 substantially improved the model performances on validation data and the trained model generalized well on the imbalanced unseen data. We trained the model for 10 epochs[1] with early stopping if the validation loss did not improve for 3 epochs. In addition, we reduced the learning rate by a factor of 0.1 if the validation loss did not improve for 2 consecutive epochs. The minimum learning rate was initialized to be 10^{-8}. The model with the lowest validation loss was returned to be used for inference.

5.3.6 Inference

The test dataset made available by the CLPSych 2019 shared task was used at inference for both the flagged/not flagged and urgent/not urgent tasks. During inference, the trained model is submitted with unseen test data to evaluate how well the model

[1]With extensive experiments, we identified 10 epochs as the optimal number of epochs for the model to be trained, considering the factors such as training loss, validation loss and variance.

has generalized. The class distribution of the test dataset is imbalanced, unlike the training dataset, which was made balanced by using the control data provided by the task organizers. The class distribution for the flagged/not flagged task is 93/32 (i.e., approximately 74% for the positive class and 26% for the negative), and for the urgent/not urgent task, it is 80/45 (i.e., approximately 64% for positive class and 36% for the negative). With each stratified split, we calculated the macro Precision, Recall and F1 score. In addition, we calculated the macro averaged ROC AUC score and the accuracy as evaluation metrics to better understand the trained model's performances. Comparing the evaluation results obtained during each stratified split, we identified a certain level of variance which could happen due to the stochastic nature of the algorithms. Hypothetically several other factors could have contributed to the variance of the results. One key factor could be the statistical noise in the dataset (Brownlee, 2018) and especially when the data is automatically annotated. To overcome the variance in the results, we used the model averaging ensemble approach (Brownlee, 2018) (see section 2.1.3). We used the expert annotated UMD dataset to demonstrate further how well the model generalizes on unseen data. The dataset contained 245 users with "No", "Low", "Moderate", and "Severe" risk categories containing: 36, 50, 115 and 44 users. The class distribution for the flagged/not flagged task is 209/36 (i.e., approximately 85% for positive class and 15% for the negative), and for the urgent/not urgent task, it is 159/86 (i.e., approximately 65% for positive class and 35% for the negative). Even though a certain level of class imbalance can be identified, especially from the flagged/not flagged task, the prediction results on the test dataset annotated by the experts show that the trained model is well generalized.

5.4 Results

The results are measured using the macro F1 score for each task, identified using the proposed multi-task learning architecture. In addition, we have reported the macro precision and recall, accuracy, and macro averaged ROC AUC score. We used stratified shuffle splits as our evaluation protocol and used a model averaging ensemble to generate the predictions on the test data. We calculated the baseline for each task and mentioned them alongside the multi-task learning results for comparison. The baselines were calculated for each of the experiment categories (i.e., flagged/not flagged and urgent/not urgent), for each disorder (i.e., eight disorders) and for each type of disorder whether the user is diagnosed with a single mental disorder or more than one disorder in addition to their primary diagnosis. Along with the mental disorder identified with the best macro F1 score, we have mentioned the results from the experiments using the pre-trained embeddings and the auxiliary inputs. We did not conduct the experiment using the pre-trained embeddings or the auxiliary inputs with all the mental disorders but only with the disorder that produced the best performances based on the F1 score. The multi-tasks with the best results are highlighted, and the best F1 score is stated in bold font.

The results tables mentioned in the following sections consist of the columns: Multi-task (the MTL experiment), Ps (Precision for suicide ideation detection), Pm(Precision for mental illness detection), Rs (Recall for suicide ideation detection), Rm (Recall for mental illness detection), F1s (F1 score for suicide ideation detection), F1m (F1 score for mental illness detection), ACCs (Accuracy for suicide ideation detection), ACCm (Accuracy for mental illness detection), AUCs (Area under the ROC curve for suicide ideation detection), AUCm (Area under the ROC curve for mental illness

detection).

5.4.1 Task: Flagged/not Flagged

Table 5.3 demonstrates the results obtained for the task flagged/not flagged with users who self-declared a single mental disorder. Each row represents the results obtained for the two tasks within the multi-task learning environment followed by the respective baselines. For example, the "suicide + adhd" represents the multi-task learning results where the two tasks are detecting users with suicide ideation and users diagnosed with a single mental disorder. Following the results from the two tasks, we mention the baseline associated with the individual tasks (e.g., adhd baseline). The last row of the table states the suicide ideation detection baseline predicted on the test data using the multi-channel CNN architecture.

Multi-task	**Ps**	**Pm**	**Rs**	**Rm**	**F1s**	**F1m**	**ACCs(%)**	**ACCm(%)**	**AUCs**	**AUCm**
suicide + adhd	0.737	0.726	0.722	0.726	0.728	0.726	80.00	79.20	0.814	0.814
adhd baseline	-	0.686	-	0.730	-	0.694	-	73.60	-	0.837
suicide + anxiety	0.780	0.780	0.774	0.774	0.777	0.777	83.20	83.20	0.860	0.860
anxiety baseline	-	0.776	-	0.845	-	0.789	-	81.60	-	0.902
suicide + autism	0.742	0.742	0.691	0.691	0.708	0.708	80.00	80.00	0.829	0.827
autism baseline	-	0.732	-	0.757	-	0.742	-	79.20	-	0.831
suicide + bipolar	0.778	0.778	0.805	0.805	0.789	0.789	83.20	83.20	0.889	0.889
bipolar baseline	-	0.799	-	0.846	-	0.816	-	84.80	-	0.920
suicide + depress	0.795	0.795	0.769	0.769	0.780	0.780	84.00	84.00	0.880	0.880
depress baseline	-	0.730	-	0.792	-	0.738	-	76.80	-	0.876
suicide + ocd	0.790	0.785	0.743	0.753	0.761	0.767	83.20	83.20	0.872	0.871
ocd baseline	-	0.781	-	0.825	-	0.796	-	83.20	-	0.895
suicide + ptsd	0.831	0.831	0.831	0.831	**0.831**	**0.831**	87.20	87.20	0.944	0.946
ptsd baseline	-	0.828	-	0.893	-	0.848	-	87.20	-	0.946
suicide + sch	0.820	0.829	0.826	0.842	0.823	0.835	86.40	87.20	0.874	0.873
Schizophrenia baseline	-	0.776	-	0.845	-	0.789	-	81.60	-	0.885
Suicide baseline	0.647	-	0.673	-	0.654	-	71.20	-	0.726	-

Table 5.3: Task: flagged/not flagged with users diagnosed with a single mental disorder

Table 5.4 demonstrates the results obtained for the task flagged/not flagged with users who self-declared multiple mental disorders. The only difference compared to the results mentioned in Table 5.3 is that instead of using users who were diagnosed

with a single mental illness, we have used the data that contained users who were diagnosed with multiple mental disorders in addition to their primary diagnosis. After identifying the primary diagnosis that produced the best results, we combined the same diagnosis with suicide ideation data to train separate models with pre-trained word embeddings and auxiliary inputs (i.e., EMPATH categories identified through exploratory analysis). For example, after identifying that users diagnosed with PTSD and one or more other mental disorders combined with the suicide ideation detection task produced better results, we used the same inputs but mapped according to pre-trained word embeddings to train a separate model. Also, we used the EMPATH categories as auxiliary inputs to train additional models to identify their impact on model generalization. Similar to Table 5.3, the baseline predictions for each mental disorder (i.e., primary diagnosis with one or more other mental disorders) and suicide ideation detection were also reported.

Multi-task	**Ps**	**Pm**	**Rs**	**Rm**	**F1s**	**F1m**	**ACCs(%)**	**ACCm(%)**	**AUCs**	**AUCm**
suicide + adhd	0.749	0.749	0.768	0.768	0.757	0.757	80.80	80.80	0.874	0.875
adhd baseline	-	0.783	-	0.835	-	0.800	-	83.20	-	0.916
suicide + anxiety	0.785	0.785	0.753	0.753	0.767	0.767	83.20	83.20	0.892	0.893
anxiety baseline	-	0.799	-	0.846	-	0.816	-	84.80	-	0.930
suicide + autism	0.778	0.778	0.784	0.784	0.781	0.781	83.20	83.20	0.876	0.877
autism baseline	-	0.764	-	0.834	-	0.774	-	80.00	-	0.882
suicide + bipolar	0.872	0.864	0.848	0.832	0.859	0.846	89.60	88.80	0.940	0.939
bipolar baseline	-	0.801	-	0.856	-	0.819	-	84.80	-	0.925
suicide + depress	0.758	0.748	0.753	0.758	0.755	0.752	81.60	80.80	0.870	0.871
depress baseline	-	0.789	-	0.856	-	0.805	-	83.20	-	0.904
suicide + ocd	0.768	0.778	0.789	0.794	0.777	0.785	82.40	83.20	0.888	0.889
ocd baseline	-	0.765	-	0.814	-	0.780	-	81.60	-	0.906
suicide + ptsd	0.896	0.896	0.843	0.843	0.865	0.865	90.40	90.40	0.967	0.966
suicide + ptsd + EMPATH	0.915	0.915	0.848	0.848	**0.875**	**0.875**	91.20	91.20	0.952	0.951
suicide + ptsd + fastText	0.872	0.872	0.848	0.848	0.859	0.859	89.60	89.60	0.958	0.959
ptsd baseline	-	0.836	-	0.868	-	0.849	-	88.00	-	0.942
suicide + sch	0.778	0.778	0.784	0.784	0.781	0.781	83.20	83.20	0.895	0.889
Schizophrenia baseline	-	0.743	-	0.818	-	0.739	-	76.00	-	0.922
Suicide baseline	0.647	-	0.673	-	0.654	-	71.20	-	0.726	-

Table 5.4: Task: flagged/not flagged with users diagnosed with multiple mental disorders

Table 5.5 demonstrates the results obtained for the task flagged/not flagged with

users whom self-declared PTSD and one or more other mental disorders and evaluated on the UMD expert annotated data. We used the expert annotated dataset as an additional test dataset to demonstrate how well the model has trained on the proposed architecture. We evaluated the model performances using the same metrics as mentioned in the previous tables. We created the baseline to understand how well the expert annotated data generalizes on a model trained using only the UMD crowdsourced data.

Multi-task	Ps	Pm	Rs	Rm	F1s	F1m	ACCs(%)	ACCm(%)	AUCs	AUCm
suicide + ptsd	0.843	0.843	0.851	0.851	**0.847**	**0.847**	92.245	92.245	0.955	0.955
Suicide baseline (expert)	0.587	-	0.652	-	0.585	-	70.204	-	0.729	-

Table 5.5: Results obtained for the task flagged/not flagged with users diagnosed with PTSD and one or more additional mental disorders and evaluated on the expert annotated data.

5.4.2 Task: Urgent/not Urgent

Table 5.6 demonstrates the results obtained for the task urgent/not urgent with users who self-declared a single mental disorder. Similar to Table 5.3, we have randomly selected users from the SMHD dataset who were diagnosed with a single mental disorder for the mental disorder detection task. Following each of the multi-task learning outcomes, the predicted baseline is included to understand better the impact each mental disorder has on the multi-task predictions. Because the best performances were reported using users diagnosed with a single mental disorder (i.e., PTSD), we used pre-trained fastText embeddings and EMPATH categories as auxiliary inputs to train two additional models to identify if such additions to the proposed architecture could enhance the overall performances of the model. The suicide ideation detection baseline predicted using the multi-channel CNN model is stated in the last row of the

table to better understand the impact mental disorder detection task has on predicting users with suicide ideation that requires urgent attention.

Multi-task	Ps	Pm	Rs	Rm	F1s	F1m	ACCs(%)	ACCm(%)	AUCs	AUCm
suicide + adhd	0.704	0.685	0.718	0.695	0.707	0.688	72.00	70.40	0.777	0.780
adhd baseline	-	0.752	-	0.773	-	0.753	-	76.00	-	0.841
suicide + anxiety	0.785	0.785	0.750	0.750	0.761	0.761	79.20	79.20	0.842	0.843
anxiety baseline	-	0.800	-	0.812	-	0.804	-	81.60	-	0.876
suicide + autism	0.717	0.717	0.686	0.686	0.694	0.694	73.60	73.60	0.780	0.778
autism baseline	-	0.723	-	0.742	-	0.715	-	72.00	-	0.815
suicide + bipolar	0.789	0.778	0.765	0.759	0.774	0.766	80.00	79.20	0.883	0.881
bipolar baseline	-	0.802	-	0.822	-	0.807	-	81.60	-	0.886
suicide + depress	0.818	0.805	0.773	0.767	0.787	0.779	81.60	80.80	0.868	0.867
depress baseline	-	0.792	-	0.786	-	0.789	-	80.80	-	0.895
suicide + ocd	0.722	0.722	0.712	0.712	0.716	0.716	74.40	74.40	0.855	0.853
ocd baseline	-	0.766	-	0.788	-	0.756	-	76.00	-	0.863
suicide + ptsd	0.859	0.859	0.865	0.865	**0.862**	**0.862**	87.20	87.20	0.946	0.946
suicide + ptsd + EMPATH	0.863	0.849	0.856	0.834	0.859	0.840	87.20	85.60	0.943	0.943
suicide + ptsd + fastText	0.860	0.860	0.840	0.840	0.848	0.848	86.40	86.40	0.942	0.941
ptsd baseline	-	0.839	-	0.866	-	0.843	-	84.80	-	0.932
suicide + sch	0.748	0.739	0.745	0.739	0.747	0.739	76.80	76.00	0.870	0.868
Schizophrenia baseline	-	0.808	-	0.813	-	0.810	-	82.40	-	0.897
Suicide baseline	0.616	-	0.625	-	0.616	-	63.20	-	0.680	-

Table 5.6: Task: urgent/not urgent with users self-diagnoses with a single mental disorder

Table 5.7 demonstrates the results obtained for the task urgent/not urgent with users who self-declared multiple mental disorders. Similar to table 5.6, we report the prediction outcomes of the multi-task learning model and the baselines computed for each task. Different from table 5.6, the users are diagnosed with multiple mental disorders.

Table 5.8 demonstrates the results obtained for the task urgent/not urgent with users who self-declared PTSD (i.e., the single mental disorder only) and evaluated on the expert annotated data. We used the UMD expert annotated data to demonstrate how well the model generalizes at inference. To illustrate the impact mental disorders have when predicting users with suicide ideation that requires urgent attention, we included the predictions made on the UMD expert annotated data using the single-task model trained on UMD crowdsource data.

Multi-task	Ps	Pm	Rs	Rm	F1s	F1m	ACCs(%)	ACCm(%)	AUCs	AUCm
suicide + adhd	0.777	0.777	0.738	0.738	0.750	0.750	78.40	78.40	0.869	0.869
adhd baseline	-	0.775	-	0.769	-	0.772	-	79.20	-	0.855
suicide + anxiety	0.846	0.843	0.838	0.843	0.842	0.843	85.60	85.60	0.921	0.921
anxiety baseline	-	0.841	-	0.858	-	0.847	-	85.60	-	0.912
suicide + autism	0.839	0.839	0.806	0.806	0.818	0.818	84.00	84.00	0.901	0.900
autism baseline	-	0.826	-	0.854	-	0.827	-	83.20	-	0.929
suicide + bipolar	0.846	0.858	0.786	0.808	0.802	0.824	83.20	84.80	0.901	0.901
bipolar baseline	-	0.814	-	0.838	-	0.817	-	82.40	-	0.908
suicide + depress	0.784	0.784	0.775	0.775	0.779	0.779	80.00	80.00	0.881	0.881
depress baseline	-	0.797	-	0.820	-	0.801	-	80.80	-	0.893
suicide + ocd	0.760	0.776	0.742	0.764	0.748	0.769	77.60	79.20	0.867	0.866
ocd baseline	-	0.765	-	0.768	-	0.766	-	78.40	-	0.861
suicide + ptsd	0.851	0.851	0.854	0.854	**0.853**	**0.853**	86.40	86.40	0.942	0.942
ptsd baseline	-	0.841	-	0.858	-	0.847	-	85.60	-	0.938
suicide + sch	0.833	0.842	0.842	0.848	0.837	0.845	84.80	85.60	0.910	0.911
Schizophrenia baseline	-	0.823	-	0.849	-	0.826	-	83.20	-	0.937
Suicide baseline	0.616	-	0.625	-	0.616	-	63.20	-	0.680	-

Table 5.7: Task: urgent/not urgent with users self-diagnoses with multiple mental disorders

Multi-task	Ps	Pm	Rs	Rm	F1s	F1m	ACCs(%)	ACCm(%)	AUCs	AUCm
suicide + ptsd	0.851	0.851	0.839	0.839	**0.845**	**0.845**	86.12	86.12	0.924	0.922
Suicide baseline (expert)	0.639	-	0.646	-	0.641	-	66.53	-	0.643	-

Table 5.8: Results obtained for the task urgent/not urgent with users diagnosed with PTSD (i.e., the single mental disorder only) and evaluated on the expert annotated data.

5.5 Discussion

The following sections will further analyze the results obtained for the different multi-task learning experiments. The experiments were conducted as binary classification tasks using eight different mental disorders combined with users having suicide ideation. The performance of each model is assessed using the macro F1 score, and in addition, further discussions will be conducted based on Area under the receiver operating characteristic curve (ROC AUC). Evaluating the AUC scores provides valuable insight into an accurate prediction of users with suicide ideation or a mental disorder, given the false positive rate. The false positive rate indicates how many users were predicted as having suicide ideation or a mental disorder when they were not identified with

either of the conditions. A higher AUC score will indicate the prediction reliability by distinguishing users with either suicide ideation or a mental disorder from the control group. In addition to the F1 and AUC scores, we used several other metrics to evaluate the trained model on unseen data. We used the following acronyms to indicate the evaluation metrics used to measure the trained model's performances.

- Ps/Pm: Precision for suicide ideation detection/mental disorder detection.
- Rs/Rm: Recall for suicide ideation detection/mental disorder detection.
- F1s/F1m: F1 score for suicide ideation detection/mental disorder detection.
- ACCs/ACCm: Accuracy for suicide ideation detection/mental disorder detection.
- AUCs/AUCm: Area under the ROC curve for suicide ideation detection/mental disorder detection.

Apart from the accuracy score, the rest of the metrics are based on macro averaging. To further evaluate the model performances, we include the ROC curves and the precision-recall curves. The ROC curve plots the false positive rate (x-axis) and true positive rate (y-axis) over different classification thresholds. The ROC curve will provide an overview of the model reliability with respect to its generalizability. Given the data imbalance in the test data, we used the precision-recall curve to plot recall (x-axis) against precision (y-axis) over different probability thresholds to better understand model performances. To conduct an in-depth analysis of the prediction results, we use the classification report that lists the classification metrics such as precision, recall, F1 score and accuracy. The metrics for each class are summarized using macro and weighted average methods. In addition, we show the confusion matrix

to illustrate the actual and predicted class counts according to the number of, True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN).

5.5.1 Task: Flagged/not Flagged

According to Table 5.3 and 5.4, and when comparing the suicide ideation detection baseline F1 score (i.e., F1s) with the highlighted best F1 scores (i.e., for the task of predicting users with suicide ideation), we could identify that the proposed multi-task learning with mixed-parameter sharing has given significantly better results over the baseline that uses a multi-channel CNN model with single task learning. Overall, the results obtained for each of the multi-task learning experiments (e.g., "suicide + adhd", "suicide + anxiety) have produced better F1 scores compared to the suicide ideation detection baseline. When comparing the best AUCs score against suicide baseline AUCs score, it could be derived that the proposed architecture has produced a low false positive rate and a higher true positive rate to generate a strong AUC score. The best performances concerning the F1 scores (i.e., F1s and F1m) are from the tasks suicide ideation detection and PTSD mental disorder detection ("suicide + ptsd"). The tasks "suicide + ptsd" have reported better F1 scores when predicting users diagnosed with a single mental disorder (i.e., PTSD only) and multiple mental disorders (i.e., PTSD with one or more other mental disorders). Using the posts from the users diagnosed only with PTSD, the predicted F1 scores for suicide ideation and mental disorder detection are 0.831. In the same order, the AUC scores reported for the two tasks were 0.944 and 0.946. In comparison to the results obtained when users were diagnosed only with PTSD, users diagnosed with multiple mental disorders in addition to the primary diagnosis (i.e., PTSD) have produced considerably better results. According to Table

5.4, both F1s and F1m have reported 0.865 with the AUC scores of 0.967 (AUCs) and 0.966 (AUCm). To identify the impact auxiliary inputs have on the prediction outcome, we used several EMAPTH categories (see section 5.3.1) and identified that the model performances could be further enhanced. With the EMPATH features, the F1 scores improved from 0.865 to 0.875, which is around 1% of a performance gain. Given the highest F1 score using the Reddit posts of users diagnosed with PTSD (i.e., for users with PTSD and one or more other mental disorders), we used pre-trained fastText embeddings to identify the level of impact it could have over randomly initialized embeddings. Even allowing the pre-trained embeddings to be further trained on the new data, the trained model could not generalize well on the unseen data compared to when using the randomly initialized word embeddings. With pre-trained fastText embeddings, the model achieved an F1 score of 0.859 for suicide ideation and mental disorder detection tasks. Even though the performances could not surpass the highest F1 score achieved using the EMPATH categories, the model using the pre-trained word embeddings generalized well on unseen data to achieve AUC scores, 0.958 for AUCs and 0.959 for AUCm. The ROC curves 5.4(a) and 5.4(b) mentioned below demonstrate the false positive rate (x-axis) and the true positive rate (y-axis) plotted against different thresholds for each of the tasks (i.e., according to the model that reported the best F1 score). For comparison, the ROC curves 5.4(c) and 5.4(d) demonstrates the results obtained from the suicide ideation and PTSD mental disorder detection baseline models.

Figures 5.5(a) and 5.5(b) plot the precision-recall curve for the positive class over different classification thresholds given the best-performing model. The two plots demonstrate the precision-recall curve for the two tasks, suicide ideation detection

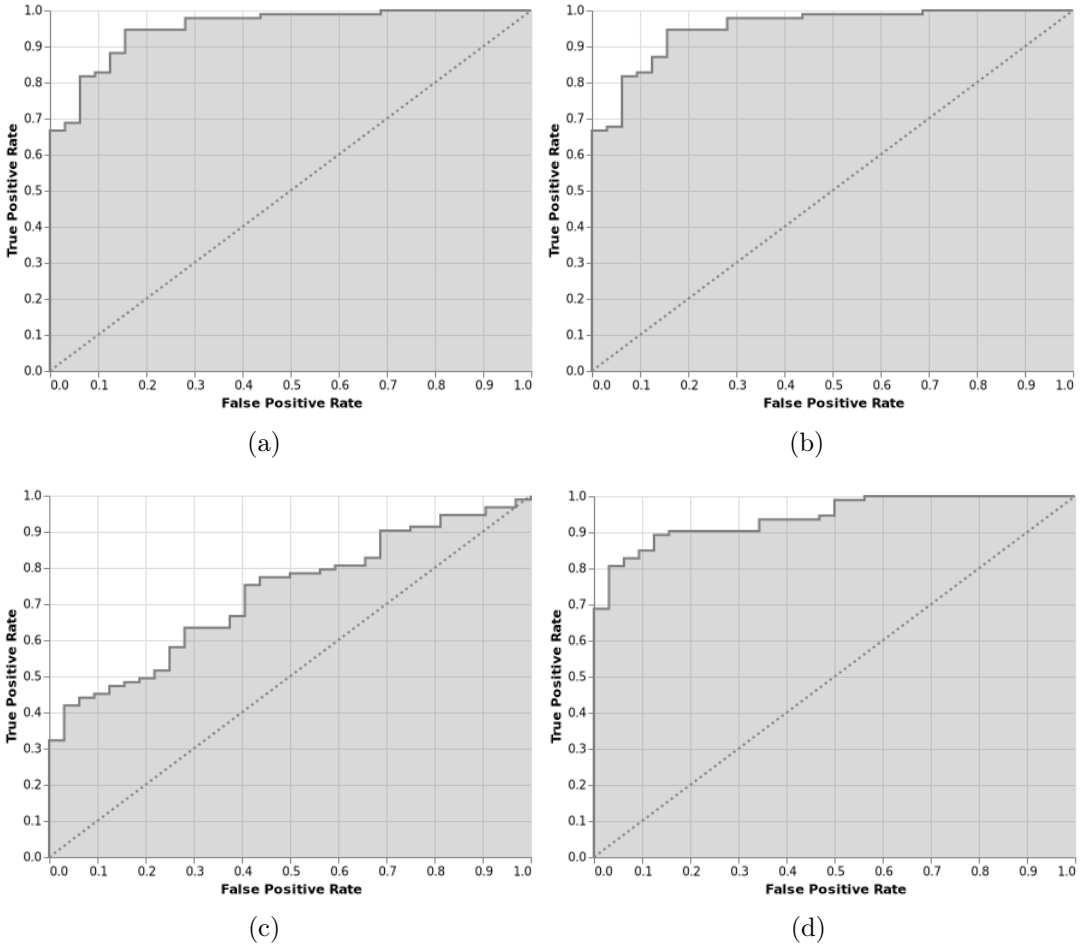
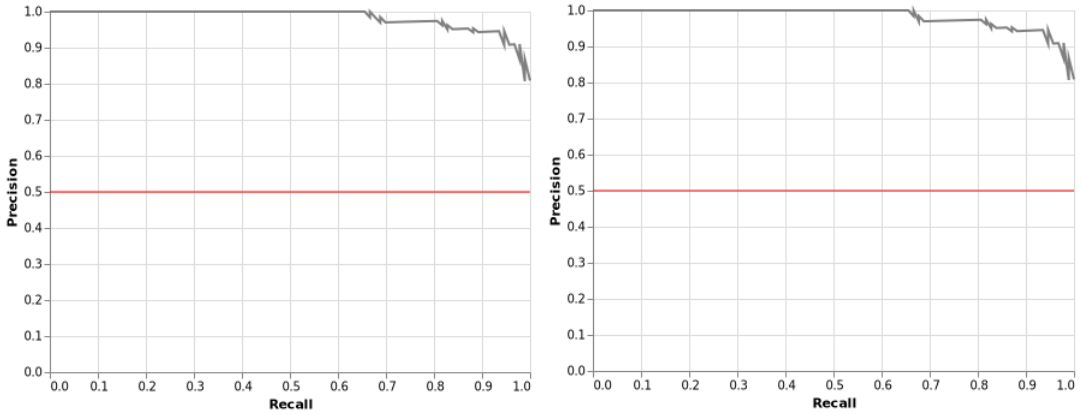


Figure 5.4: ROC curves for: (a) suicide ideation detection(suicide + ptsd + EMPATH) (b) PTSD mental disorder detection(suicide + ptsd + EMPATH) (c) Suicide baseline (d) ptsd baseline

and PTSD mental disorder detection. Evaluating the two plots, we could identify that the trained model performs significantly better than a random baseline and, given the positive class, both precision and recall scores can be above 0.90.

To further demonstrate the model performance, we generated the classification report (Pedregosa et al., 2011) to validate the individual class (i.e., positive class represented with "1" and the control group with "0") contribution towards the model's



(a) (b)

Figure 5.5: precision-recall curves for: (a) suicide ideation detection(suicide + ptsd + EMPATH) (b) PTSD mental disorder detection(suicide + ptsd + EMPATH)

overall performance. Table 5.10 lists both macro and weighted precision, recall, F1 score in addition to overall accuracy, and the class distribution (i.e., Users), given the test data. We did not include the mental illness detection task classification report as the mental illness, and suicide ideation detection results are pretty similar. Getting similar results for both tasks is highly probable given the proposed multi-task learning architecture, where both the tasks have to perform equally well to reduce the overall loss.

	Precision	**Recall**	**F1 score**	**Users**
Class 0	0.92	0.72	0.81	32
Class 1	0.91	0.98	0.94	93
Macro average	0.92	0.85	**0.88**	125
Weighted average	0.91	0.91	0.91	125
Accuracy		0.91		125

Table 5.9: Classification report with the best results for suicide ideation detection obtained from the task "suicide + ptsd" with EMPATH categories as auxiliary inputs

When evaluating the classification report, table 5.9, we could see that the positive

class (i.e., users with suicide ideation) has produced a better F1 score than the control class. The control class results have negatively impacted the overall performances of the model. Analyzing the confusion matrix mentioned in table 5.10, we could see that given a small number of data points (n=125) and 32 instances in the control class, only 23 (TP) have been correctly classified where 9 (FN) instances were wrongly classified as negatives while 2 (FP) instances were wrongly classified as positives. The number of false negatives (FN) indicates that from the 32 control users, 9 users were predicted to have suicidal thoughts, and 2 users who were having suicidal thoughts were predicted as neurotypicals. Concerning the low recall in the control class, we could assume that because having similar content to "Low" risk users might have categorized "No" risk users into users with suicide ideation. The same outcome can be combined with the mental illness detection task where users recognized as not having a mental illness could share similar features with users having low mental illness predictors.

n=125	**Predicted: 0/1**	**Predicted: 1/0**
Actual: 0/1	23 (TN)/(TP)	9 (FP)/(FN)
Actual: 1/0	2 (FN)/(FP)	91 (TP)/(TN)

Table 5.10: Confusion matrix with the best results for suicide ideation detection obtained from the task "suicide + ptsd" with EMPATH categories as auxiliary inputs

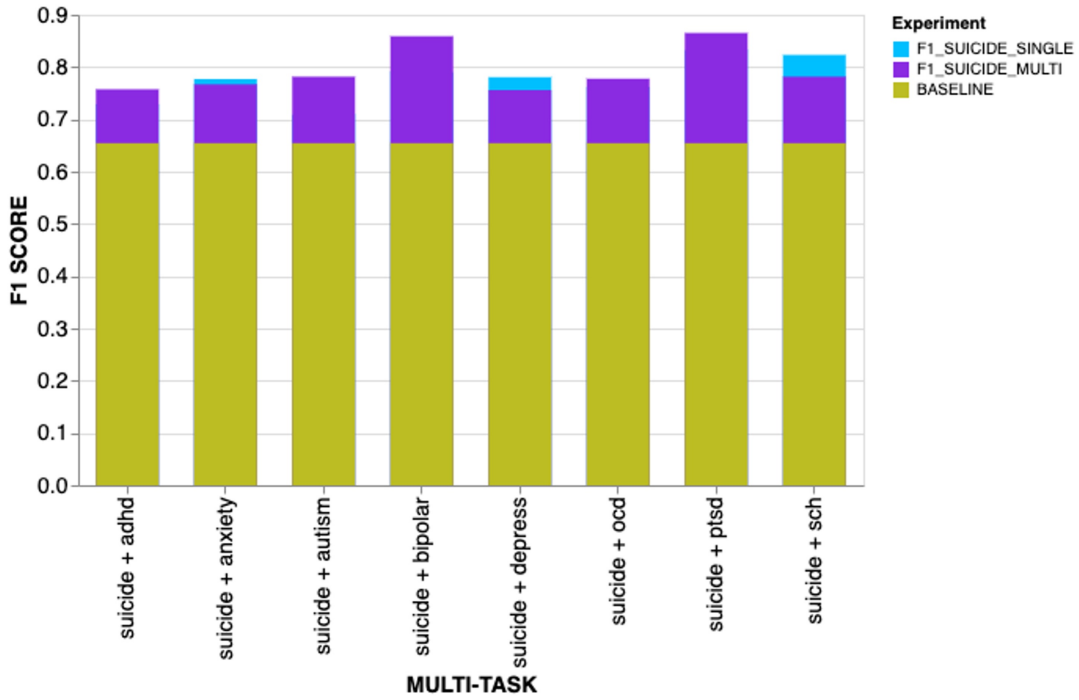
To analyze the overall outcome of our experiments in detecting suicide ideation and mental disorders using the proposed MTL with mixed parameter sharing, we have included Figures 5.6, 5.7, 5.13 and 5.14 that demonstrate the combination of tasks in the x-axis with their associated macro F1 scores in the y-axis. The F1 scores are from the tasks suicide ideation and mental illness detection without any additional

experiments. For example, when predicting users with suicide ideation and PTSD, we did not use the best results obtained by combining the auxiliary inputs and instead, we used results attained when using the data only from the two tasks. The outcome of each experiment is represented in the layered bar chart, where the labels mentioned in the chart legend indicate the F1 score for the related experiment:

- F1_SUICIDE_SINGLE: F1 score for suicide ideation detection task given the users with a single mental disorder.
- F1_SUICIDE_MULTI: F1 score for suicide ideation detection task given the users with multiple mental disorders.
- BASELINE: Baseline F1 score for suicide ideation detection.
- F1_MENTAL_SINGLE: F1 score for mental illness detection task given the users with a single mental disorder.
- F1_MENTAL_MULTI: F1 score for mental illness detection task given the users with multiple mental disorders.
- BASELINE_SINGLE: Baseline F1 score for mental illness detection given the users with a single mental disorder.
- BASELINE_MULTI: Baseline F1 score for mental illness detection given the users with multiple mental disorders.

According to figure 5.6, we could see that users diagnosed with multiple mental disorders (i.e., apart from the user's primary diagnosis being anxiety, depression and schizophrenia) share more hidden features with users having suicide ideation

than with users diagnosed with a single mental disorder. In general, it is clear that, given the particular dataset, mental disorders have positively impacted the suicide ideation detection task in the MTL environment with mixed parameter sharing. Also, considering the eight mental disorders, different mental disorders have imposed a distinctive impact on suicide ideation detection. For example, users diagnosed with either PTSD, bipolar or schizophrenia tend to have shared more hidden features with users identified with suicidal thoughts than users with other mental disorders. All three mental disorders mentioned before have produced F1 scores greater than 0.80 (i.e., for both F1s and F1m) with AUC scores between 0.88 and 0.97.



	suicide + adhd	suicide + anxiety	suicide + autism	suicide + bipolar	suicide + depress	suicide + ocd	suicide + ptsd	suicide + sch
F1_MENTAL_SINGLE	0.728	0.777	0.708	0.789	0.78	0.761	0.831	0.823
F1_MENTAL_MULTI	0.757	0.767	0.781	0.859	0.755	0.777	0.865	0.781
BASELINE	0.654	0.654	0.654	0.654	0.654	0.654	0.654	0.654

Figure 5.6: Overall results for the suicide ideation detection task with users diagnosed with single or multiple mental disorders. It also includes the suicide ideation detection baseline.

In addition to identifying the impact mental disorders have on suicide ideation detection, we analyzed the impact suicide ideation detection task has on mental disorder detection. When computing the baseline for suicide ideation detection, we did not use a majority class baseline and instead used the same underlying multi-task learning architecture but for single task learning. The multi-channel CNN network (i.e., with the same configuration of hyperparameters) used for task-specific learning within the proposed multi-task learning architecture was used for single task learning. When comparing our proposed baseline with the majority class baseline accuracy, it could be identified that, apart from a few occurrences, the proposed baseline model has produced better accuracies, especially for the mental disorder detection task. The majority class baseline accuracies for mental illness and suicide ideation detection are 74.4% (i.e., for flagged/not flagged) and 64% (i.e., urgent/not urgent). For example, in the flagged/not flagged task, the proposed baseline accuracy when predicting users with PTSD is 87.3%, which is considerably higher than the majority class baseline of 74.4%. Apart from when detecting ADHD (i.e., for single mental disorder) within the flagged/not flagged task, which has generated a slightly lower baseline accuracy (i.e., 73.6%), the rest of the mental disorders have generated a baseline accuracy considerably higher than the majority class baseline (i.e., for single and multiple mental illness detection within both flagged/not flagged and urgent/not urgent tasks). However, the proposed baseline accuracies for suicide ideation detection are slightly less than the majority class baseline. The proposed baseline accuracies for suicide ideation detection under the flagged/not flagged and urgent/not urgent tasks are 71.2% and 63.2%, which are slightly less than the majority class baselines 74.4% and 64%. Even though the suicide ideation detection majority class baseline is slightly

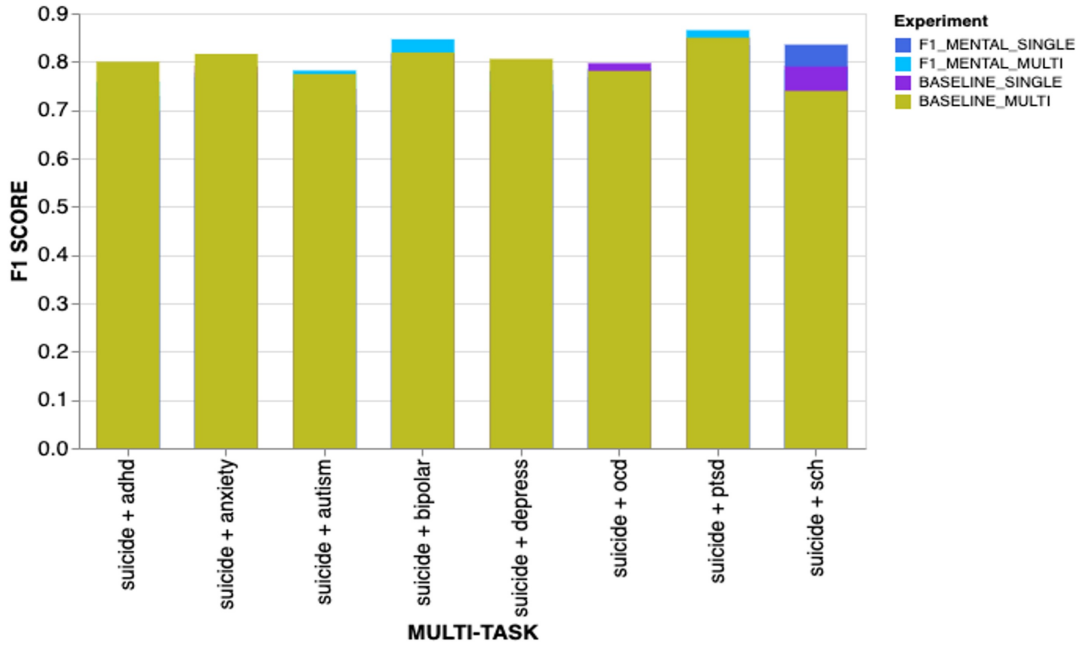
higher than the proposed baseline, the prediction accuracies of our proposed MTL model are considerably higher.

According to figure 5.7, that is when measuring the impact suicide ideation detection has on mental illness detection, we could identify that using data associated with certain mental disorders has not improved the model performances over the baseline predictions. For example, the tasks; "suicide + adhd", "suicide + anxiety", "suicide + depress" and "suicide + ocd", have not managed to improve the F1 scores (i.e., F1m) for mental illness detection. The best F1 baseline (i.e., from either single or multiple mental disorders) for ADHD, anxiety, depression, and OCD are 0.800, 0.816, 0.805, 0.796. The highest F1 score for the same mental disorders is 0.757, 0.777, 0.780, 0.785. Given the computed F1 scores and the baseline, it could be derived that the features shared between the two tasks have managed to improve only the performances of suicide ideation detection and have hindered the performances of mental illness detection. Even though the tasks were aligned, the information being transferred from one task to another has negatively impacted its performance (Wu et al., 2020). Hypothetically, users with certain mental disorders could have features shared with users with suicidal thoughts, but on the contrary, users with suicidal thoughts might not have features that could be used to further distinguish users diagnosed with certain mental disorders from neurotypicals. The argument can be further extended to state that even though mental disorders such as depression have a strong correlation with suicide ideation (Brådvik, 2018), given the SMHD dataset, the users diagnosed with depression might not have published content containing characteristics that link suicide risk with depression. For example, similar to the suicidality predictors identified among those who experienced depression, such as

"depression history and severity", "comorbid mental illness", "help seeking", and "socio demographic characteristics" (Handley et al., 2018), the neural network needs to extract certain distinctive features to discover the level of interrelatedness between suicide ideation and depression.

Apart from the aforementioned mental disorders that have not gained any advantage with respect to their detection performance given the multi-task learning environment, the remaining mental disorders, autism, bipolar, PTSD and schizophrenia, have reported improvement over the single-task baseline. Users diagnosed with multiple mental disorders in addition to their primary diagnosis, autism, bipolar, and PTSD, have reported improved F1 scores; 0.781, 0.846, 0.865 in comparison to their baseline scores; 0.774, 0.819, 0.849. Similar to the suicide ideation detection results compared in figure 5.6, the mental illness detection task for schizophrenia has reported the best F1 score when used with users diagnosed with a single mental disorder. Compared with the single-task baseline, the F1 score for detecting users with schizophrenia has improved from 0.789 to 0.835.

When further evaluating figures 5.6 and 5.7, we could see that users diagnosed with multiple mental disorders have shared more features with users having suicidal thoughts than those diagnosed with a single mental disorder. Even when it comes to predicting the baseline for mental disorders, we could clearly distinguish that users diagnosed with multiple mental disorders in addition to their primary diagnosis have produced better F1 scores compared to the users diagnosed with a single mental disorder. One of the key observations we came across is when using users diagnosed with schizophrenia, where the best F1 scores for both suicide ideation and mental illness detection were discovered when using the posts published by users diagnosed



	suicide + adhd	suicide + anxiety	suicide + autism	suicide + bipolar	suicide + depress	suicide + ocd	suicide + ptsd	suicide + sch
F1_MENTAL_SINGLE	0.726	0.777	0.708	0.789	0.78	0.767	0.831	**0.835**
F1_MENTAL_MULTI	0.757	0.767	**0.781**	**0.846**	0.752	0.785	**0.865**	0.781
BASELINE_SINGLE	0.694	0.789	0.742	0.816	0.738	**0.796**	0.848	0.789
BASELINE_MULTI	**0.800**	**0.816**	0.774	0.819	**0.805**	0.78	0.849	0.739

Figure 5.7: Overall results for the mental illness detection task with users diagnosed with single or multiple mental disorders. It also includes the mental illness detection baseline for both single and multiple mental disorders.

with a single mental disorder rather than multiple mental disorders. Even though we could not conclusively state the differences between schizophrenia and other mental disorders used in our research, it could be derived that specific psychosis characteristics unique to schizophrenia could have given the dataset its unique properties.

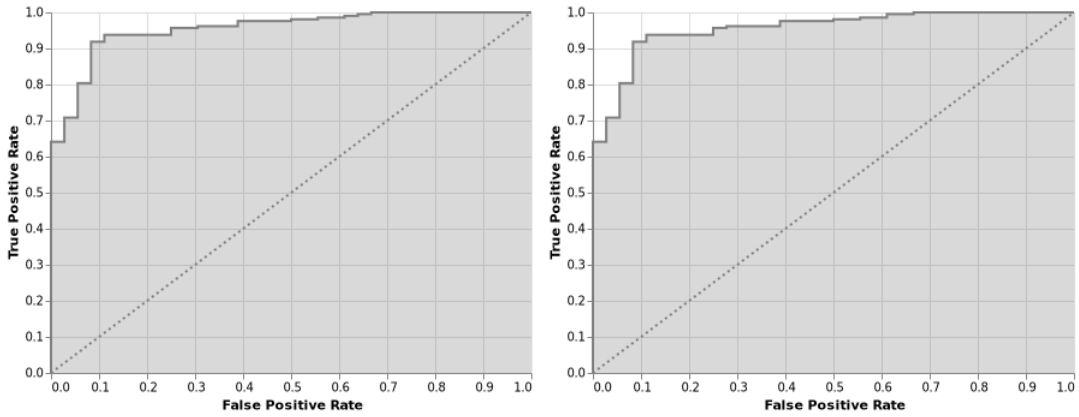
We used the expert annotated data from the UMD dataset during inference to further demonstrate how well the trained model generalizes on unseen data. Due to the limited number of instances annotated by the experts, the expert annotated data was used only for testing purposes and not during training or the validation phases. We did not use the test data on all the different models trained using eight mental

disorders, except the best performing model. The model trained on suicide ideation and PTSD (i.e., users diagnosed with one or more mental disorders in addition to PTSD) data is used to predict 245 users annotated by the experts. In line with the suicide ideation detection baseline, we created a separate baseline for suicide ideation detection using the expert annotated data. The baseline was created using expert annotated data as the test dataset, while the model was trained using crowdsourced data. According to table 5.5, we obtained an F1 score of 0.585 and an AUC score of 0.729 as the baseline. Testing on the expert annotated data, we obtained 0.847 as the F1 score for both tasks. In addition, the AUC scores for suicide ideation detection and mental disorder detection are, 0.9550 (i.e., AUCs) and 0.9553 (i.e., AUCm). According to the AUC scores, we could identify how well the model has generalized on unseen data for suicide ideation and mental illness detection, even with a different class distribution than the crowdsourced test dataset. Mentioned in figure 5.8 are the ROC curves for suicide ideation (figure 5.8(a)) and mental disorder (figure 5.8(b)) detection tasks when using the expert annotated data as the test data.

According to figure 5.9, the precision-recall curve performs better than the random baseline, and for both the tasks, we could identify that precision and recall scores report a value greater than 0.90 (i.e., for the positive class).

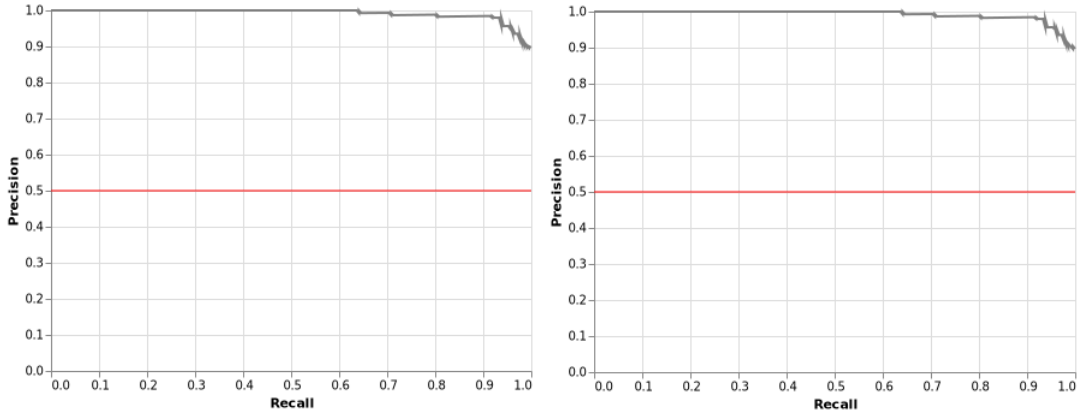
When further analyzing the prediction outcome on the expert annotated data using the classification report 5.11, it is clear that similar to when using the crowdsourced data, the precision and recall scores reported on the control class are less than those computed on the positive class.

Looking at table 5.11, we could see that the control class negatively impacts the overall evaluation metrics. For example, the macro F1 score in the positive class is



(a) (b)

Figure 5.8: ROC curves for: (a) suicide ideation detection(suicide + ptsd) (b) PTSD mental disorder detection(suicide + ptsd). Using the expert annotated data as the test data.



(a) (b)

Figure 5.9: precision-recall curves for: (a) suicide ideation detection(suicide + ptsd) (b) PTSD mental disorder detection(suicide + ptsd). Using the expert annotated data as the test data.

	Precision	Recall	F1 score	Users
Class 0	0.73	0.75	0.74	36
Class 1	0.96	0.95	0.95	209
Macro average	0.84	0.85	**0.85**	245
Weighted average	0.92	0.92	0.92	245
Accuracy		0.92		245

Table 5.11: Classification report with the results for suicide ideation detection using the expert annotated data

0.95, but for the control class, it is 0.74. According to the confusion matrix in table 5.12, and given the control class, both the precision and recall have been impacted where a corresponding number of instances were misclassified as false negatives and false positives. One key reason for the misclassification could be the class distribution where the expert annotated test data is imbalanced (i.e., approximately 85% positive class and 15% negative) than the crowdsource annotated test data (i.e., approximately 74% positive class and 26% negative). If applied in a clinical setting, having control users classified as suicidal can be considered as having less of an impact than classifying users with suicidal thoughts as control users. We will conduct further experiments to identify the possibilities of reducing false positives and false negatives in our future work. Similar results were identified in the classification report and the confusion matrix of the mental illness detection task where an equal number of users to that of the expert annotated data were randomly selected during inference.

n=245	**Predicted: 0/1**	**Predicted: 1/0**
Actual: 0/1	27 (TN)/(TP)	9 (FP)/(FN)
Actual: 1/0	10 (FN)/(FP)	199 (TP)/(TN)

Table 5.12: Confusion matrix with the results for suicide ideation detection using the expert annotated data

5.5.2 Task: Urgent/not Urgent

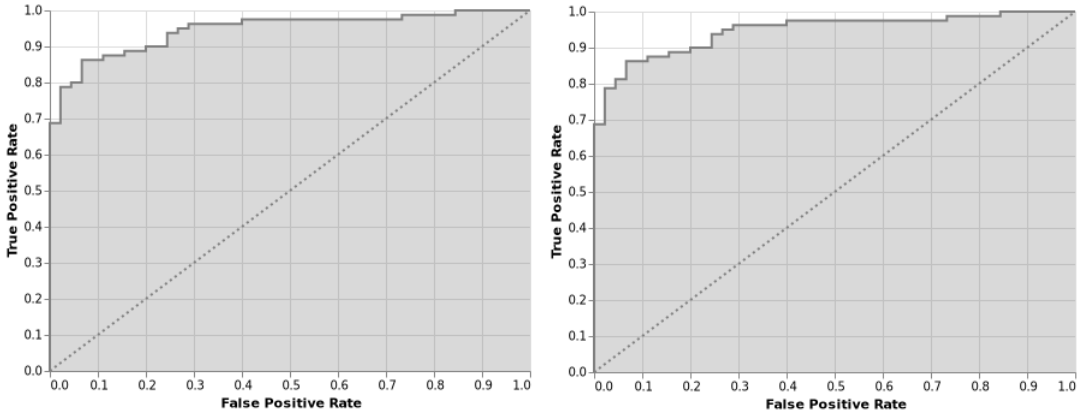
We conducted a similar set of experiments as previously mentioned in the task flagged/not flagged for the task urgent/not urgent. According to Tables 5.6 and 5.7, we could identify that using the proposed multi-task learning with mixed parameter sharing model has produced significantly better results for the suicide ideation detection task for users that requires urgent attention. The F1 scores for both suicide ideation detection and PTSD mental disorder detection, when used with data from users who self-declared a single mental disorder (i.e., only PTSD), was 0.862, which is more than 20% of an increase in comparison to the suicide ideation detection baseline that is 0.616. The proposed architecture has produced significantly better results than the majority class baseline (i.e., 0.64) for suicide ideation detection. Similar to the flagged/not flagged task findings, we did not find significant improvement of performances on mental illness detection (i.e., either single or multiple) compared to the baseline results of each mental disorder. For example, when considering users diagnosed only with PTSD, the performance gain with respect to the macro F1 score is nearly 2%, where the F1 score with multi-task learning is 0.862 compared to the baseline F1 0.843. A similar pattern in the results was identified with suicide ideation and mental disorder detection AUC scores, where for the suicide ideation detection task, an improvement of around 26% is identified while for mental disorder detection, it is a bit more than 1%. The reported baseline AUC scores for suicide ideation and PTSD mental illness detection are 0.680 and 0.932, while the reported AUC scores for the proposed architecture are 0.9463 (AUCs) and 0.9466 (AUCm). When considering the reported metrics, we could identify that much of the knowledge being transferred is from the mental illness detection task to the suicide ideation detection task, confirming

the impact mental disorders have on users with suicidal thoughts who requires urgent attention.

Similar to the experiments conducted under the task flagged/not flagged, we used EMPATH auxiliary inputs and fastText embeddings to train two different models to identify the prospects of improving the prediction results on unseen data. Using the EMPATH categories, we could not increase the performance over the model trained without any auxiliary inputs. Given the two models, the one provided with the EMPATH categories managed to record an F1 score of 0.859, and with the pre-trained fastText embeddings, it managed to achieve an F1 score of 0.848. For suicide ideation and mental illness detection, the proposed architecture has achieved AUC scores of 0.9463 and 0.9466, which can be considered reliable outcomes, especially when predicting a low false positive rate for users with suicidal thoughts and PTSD (i.e., single mental disorder). Figure 5.10 state the ROC curves for the best-generalized model trained on suicide ideation (figure 5.10(a)) and PTSD mental illness detection (figure 5.10(b)) tasks.

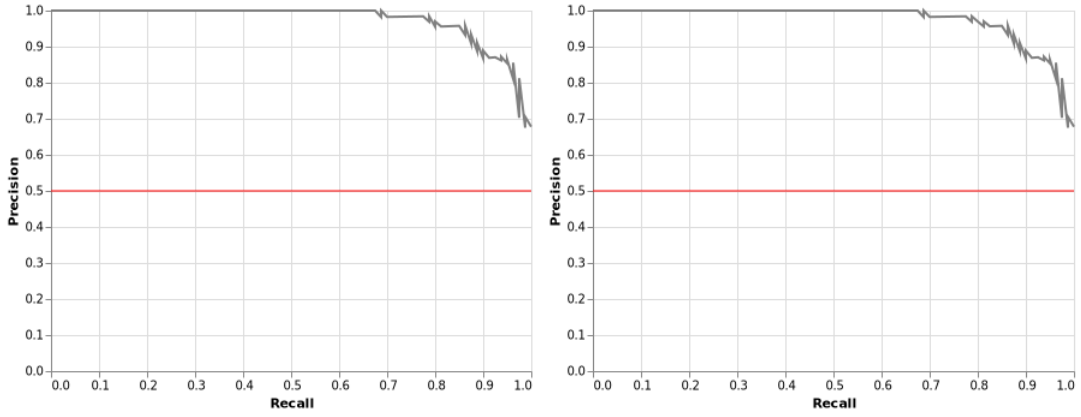
Given the imbalanced nature of the test dataset, we created the precision-recall plots (figure 5.11) for both the tasks (i.e., predicting suicide ideation and PTSD only) to evaluate the performances of the trained model over different classification thresholds.

With similar results, we could identify that given the positive class, both suicide ideation and mental illness detection have achieved precision scores above 0.9, with a recall closer to 0.9. In general, both the tasks have achieved precision and recall scores above the random baseline for all the classification thresholds being tested. It is clear how well the model has generalized when considering both AUC and precision-recall



(a) (b)

Figure 5.10: ROC curves for: (a) suicide ideation detection(suicide + ptsd) (b) PTSD mental disorder detection(suicide + ptsd)



(a) (b)

Figure 5.11: precision-recall curves for: (a) suicide ideation detection(suicide + ptsd) (b) PTSD mental disorder detection(suicide + ptsd)

curves for both the tasks. When analyzing the classification report (table 5.13) and the confusion matrix (table 5.14), we could identify that both the control class and the positive class has generated stable results, and the positive class has produced better results over the control class for both suicide ideation and mental illness detection tasks. Because the results for both the tasks are quite similar, we have mentioned only the suicide ideation detection results for detailed analysis.

	Precision	**Recall**	**F1 score**	**Users**
Class 0	0.81	0.84	0.83	45
Class 1	0.91	0.89	0.90	80
Macro average	0.86	0.87	**0.86**	125
Weighted average	0.87	0.87	0.87	125
Accuracy		0.87		125

Table 5.13: Classification report with the best results for suicide ideation detection of users requiring urgent attention obtained from the task "suicide + ptsd".

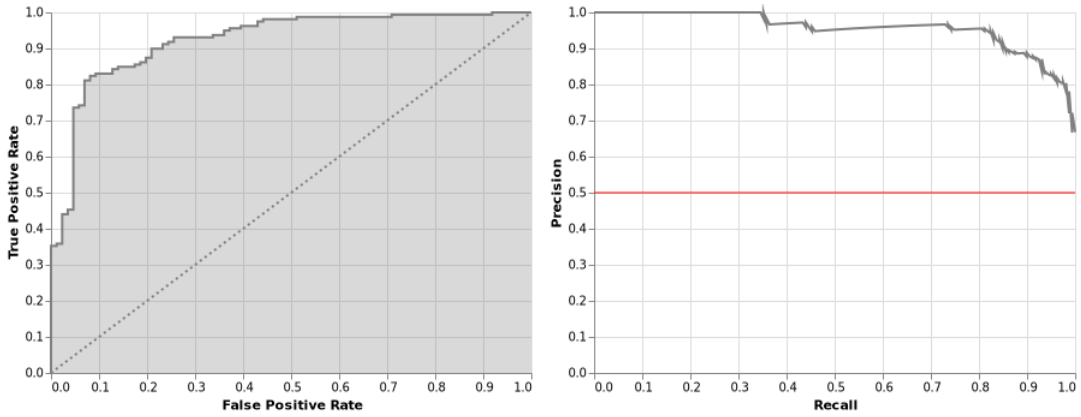
n=125	**Predicted: 0/1**	**Predicted: 1/0**
Actual: 0/1	38 (TN)/(TP)	7 (FP)/(FN)
Actual: 1/0	9 (FN)/(FP)	71 (TP)/(TN)

Table 5.14: Confusion matrix with the best results for suicide ideation detection of users that requires urgent attention obtained from the task "suicide + ptsd".

When comparing the control group results with the positive class, it could be identified that similar to the task flagged/not flagged, the control class has produced a lower macro F1 score (0.83) than the one generated by the positive class (i.e., 0.90). Further evaluating the control class metrics, we could see that 9 users identified as having suicidal thoughts were predicted as neurotypicals while 7 users labelled as neurotypicals were predicted as having suicidal thoughts. Due to the false positives and the false negatives, the control group has reported a lower F1 score (0.83) than

that of the positive class (0.90). However, we could identify that the model has generalized well given the macro F1 score of 0.86. To further demonstrate how well the trained model using our proposed architecture generalizes on unseen data, we used the UMD expert annotated data to predict users with suicide ideation and PTSD mental disorder (i.e., PTSD only). Table 5.8 shows that both suicide ideation and mental illness detection tasks have achieved an F1 score of 0.845. In addition, both the tasks have produced AUC scores of 0.924 (AUCs) and 0.922 (AUCm). Also, we obtained a baseline for single task learning (i.e., for suicide ideation detection) using the UMD expert annotated data and achieved macro F1 and AUC scores of 0.641 and 0.643, respectively. Compared to the crowdsourced test results and the calculated baseline (i.e., for suicide only) using the expert annotated data, we could see that the trained model has generalized well given the new test dataset with a different class distribution to that of the crowdsourced test data. Figure 5.12 represent the ROC curve and the precision-recall curve for the task suicide ideation detection that further demonstrates the generalizing capabilities of the trained model based on the proposed multi-task learning architecture. According to the precision-recall curve, precision and recall have performed well above the random baseline and have generated scores around 0.90 given the positive class. To further demonstrate the effectiveness of the trained model on unseen and imbalanced test data, we have included the classification report (table 5.15) and the confusion matrix (table 5.16).

According to the classification report 5.15, we could identify a similar pattern to that of the results reported for the task flagged/not flagged (table 5.11), where both the tasks have generated an F1 score closer to 0.85. Given the two tasks and their respective control class macro F1 scores 0.80 (i.e., for urgent/not urgent) and 0.74



(a) (b)

Figure 5.12: ROC(a) and precision-recall(b) curves for Suicide ideation detection(suicide + ptsd) using expert annotated data.

	Precision	Recall	F1 score	Users
Class 0	0.82	0.77	0.80	86
Class 1	0.88	0.91	0.90	159
Macro average	0.85	0.84	**0.85**	245
Weighted average	0.86	0.86	0.86	245
Accuracy		0.86		245

Table 5.15: Classification report for suicide ideation detection of the users requiring urgent attention using the expert annotated data.

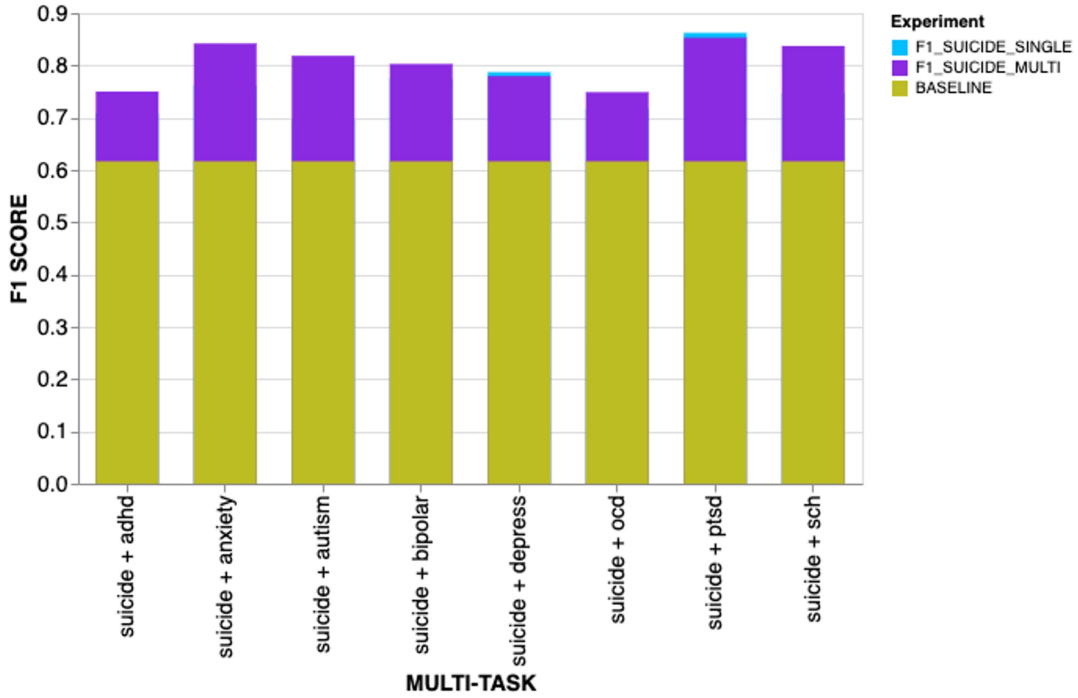
n=245	**Predicted: 0/1**	**Predicted: 1/0**
Actual: 0/1	66 (TN)/(TP)	20 (FP)/(FN)
Actual: 1/0	14 (FN)/(FP)	145 (TP)/(TN)

Table 5.16: Confusion matrix for suicide ideation detection of users that requires urgent attention using the expert annotated data.

(i.e., for flagged/not flagged), we could ascertain that the suicide risk categories "No" and "Low" share a common set of hidden features which are automatically discovered during model training.

Figure 5.13 demonstrates the overall performances generated by the proposed MTL model on predicting users with suicide ideation that requires urgent attention. The layered bar chart demonstrates the F1 scores obtained for suicide ideation detection using data from users diagnosed with either single or multiple mental disorders. In addition, the baseline results for predicting users with suicide ideation are also included for comparison. Given the suicide ideation prediction baseline F1 score of 0.616, the proposed MTL approach with mixed parameter sharing has surpassed the baseline performances with a significant margin. Also, we could see that users diagnosed only with PTSD have produced better results than the other mental disorders with an F1 score of 0.862. However, apart from PTSD and depression, users diagnosed with multiple mental disorders share more hidden features, with users having suicide ideation who requires urgent attention. Even though all the mental disorders have enhanced the performances of suicide ideation detection, different mental disorders have demonstrated a varying degree of impact on suicide ideation detection within the multi-task learning environment. For example, users diagnosed with multiple mental disorders in addition to the primary diagnosis of schizophrenia, bipolar, autism and anxiety have achieved F1 scores, 0.837, 0.802, 0.818, 0.842. Contrary to the flagged/not flagged task, users diagnosed with anxiety and autism, in addition to being diagnosed with multiple mental disorders, have contributed more towards identifying users with suicide ideation who require urgent attention.

According to tables 5.6, 5.7, and figure 5.14, we could identify that, even though



	suicide + adhd	suicide + anxiety	suicide + autism	suicide + bipolar	suicide + depress	suicide + ocd	suicide + ptsd	suicide + sch
F1_MENTAL_SINGLE	0.707	0.761	0.694	0.774	**0.787**	0.716	**0.862**	0.747
F1_MENTAL_MULTI	**0.750**	**0.842**	**0.818**	**0.802**	0.779	**0.748**	0.853	**0.837**
BASELINE	0.616	0.616	0.616	0.616	0.616	0.616	0.616	0.616

Figure 5.13: Overall results for the suicide ideation detection task(i.e., for users requiring urgent attention) with users diagnosed with single or multiple mental disorders. Also includes the suicide ideation detection baseline.

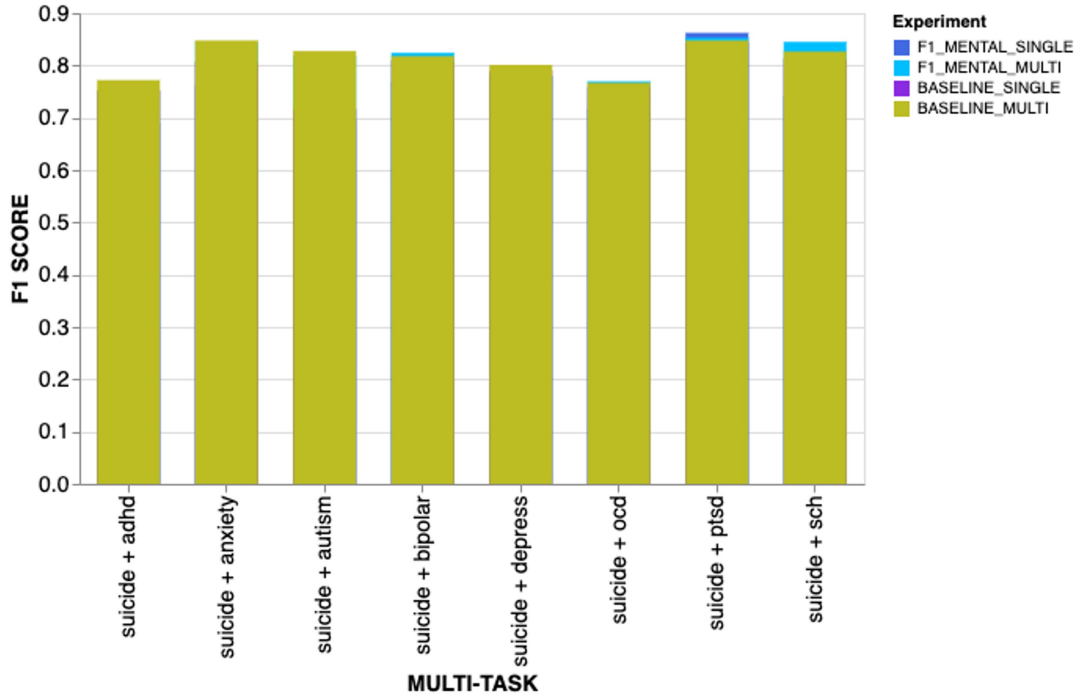
the task mental disorder detection has significantly improved suicide ideation detection, the detection of mental disorders has not considerably improved when trained alongside the suicide ideation detection task. For example, given the users diagnosed with either a single disorder or multiple mental disorders in addition to their primary diagnosis, prediction results of the disorders, ADHD, anxiety, autism and depression, did not improve over their respective baseline predictions. However, all the mental disorder prediction task accuracies have improved compared to the majority class baseline accuracy (i.e., 0.64). Given the users with multiple mental disorders in addition to their primary diagnosis of ADHD, anxiety, autism and depression, the baseline

predictions for the mental disorders are 0.772, 0.847, 0.827 and 0.801. Compared to the baseline F1 scores, the predicted outcomes from the proposed MTL model are 0.750, 0.843, 0.818 and 0.779. Given the baseline predictions, we can derive that the knowledge shared between the two tasks has benefitted suicide ideation detection more than mental illness detection. We could also identify that the information transferred from the suicide ideation detection task to the mental disorder detection task has negatively impacted mental illness predictions.

Apart from the above-mentioned mental disorders, bipolar, OCD, PTSD and schizophrenia mental disorders have shared more knowledge with suicide ideation detection. Except for PTSD, the rest of the disorders associated with multiple mental illnesses have improved over their baseline predictions. When predicting the baseline F1 scores for users diagnosed with single or multiple mental disorders, we could see that being diagnosed with multiple mental disorders in addition to their primary diagnosis have generated a more substantial baseline compared to the baseline generated using users diagnosed with a single mental disorder. Similar to the results obtained when using data from users diagnosed only with schizophrenia in the flagged/not flagged task, we could identify that users diagnosed only with PTSD had produced the best F1 score when predicting users with suicide ideation who require urgent attention.

5.6 Comparison to Related Work

We used the CLPSych 2019 test dataset provided for "Task B" without any modification for inference to directly compare our results with the state-of-the-art results generated by the task participants. Due to our research objective and the architectural requirements, we did not conduct our research on predicting the four suicide risk



	suicide + adhd	suicide + anxiety	suicide + autism	suicide + bipolar	suicide + depress	suicide + ocd	suicide + ptsd	suicide + sch
F1_MENTAL_SINGLE	0.688	0.761	0.694	0.766	0.779	0.716	**0.862**	0.739
F1_MENTAL_MULTI	0.750	0.843	0.818	0.824	0.779	**0.769**	0.853	**0.845**
BASELINE_SINGLE	0.753	0.804	0.715	0.807	0.789	0.756	0.843	0.810
BASELINE_MULTI	**0.772**	**0.847**	**0.827**	**0.817**	**0.801**	0.766	0.847	0.826

Figure 5.14: Overall results for the mental illness detection task with users diagnosed with single or multiple mental disorders. Also includes the baseline predictions for detecting users with either a single or a multiple mental disorder.

categories, except our research was focused on the flagged/not flagged and urgent/not urgent tasks. Table 5.17 demonstrates the best results obtained by the task participants on the two tasks. The CLPSych 2019 test dataset contained 125 users where for the flagged/not flagged task, 93 users were identified as having suicidal thoughts and 32 users without. For the urgent/not urgent task, 80 users were identified with suicide ideation requiring urgent attention, while 45 users were categorized into the control group. For comparison, we used our best results predicted using the content from users diagnosed with either single or comorbidity of disorders.

Submissions	F1(flagged/not flagged)	F1(urgent/not urgent)
Matero et al. (2019)	0.821	0.816
CAMH	**0.91**	0.812
Iserman et al. (2019)	0.848	0.775
Mohammadi et al. (2019)	0.843	0.718
MTL with mixed parameter sharing(using SMHD data)	0.875	**0.862**

Table 5.17: Related work comparison for suicide ideation detection(using SMHD data)

The results mentioned in table 5.17 are compared based on the macro F1 score and have included the teams that have produced the best results for both flagged/not flagged and urgent/not urgent tasks. Due to the reason that the team "CAMH" has not provided a technical paper explaining their proposed solution, we could not extract further details on how they have achieved the best results for task flagged/not flagged. In addition to the best-performing teams' results, we have included the teams ranked third for flagged/not flagged and urgent/not urgent tasks. We have highlighted our results "MTL with mixed parameter sharing" and emphasized the best results compared to the task participants.

For the flagged/not flagged task, team "CAMH" has produced the best results with an F1 score of 0.91, followed by Iserman et al. (2019) with an F1 score of 0.848. The team ranked first for detecting all four risk categories have obtained an F1 score of 0.821 for the same task. In comparison, our proposed architecture with mixed parameter sharing has generated an F1 score of 0.875, which is placed second in the overall performance. However, for the task urgent/not urgent, our proposed model has obtained an F1 score of 0.862, which is an increase closer to 5% over the state-of-the-art results obtained by Matero et al. (2019). For the same task, team "CAMH" have reported an F1 score of 0.812. When evaluating the comparisons further, we could identify that our proposed architecture has produced consistent results over both the tasks where most of the other participants, except for Matero et al. (2019), have

demonstrated a significant margin between the two results. Given these comparisons, we could identify that our proposed architecture is well generalized over both tasks.

To compare how well our proposed solution has predicted users with mental disorders, we compared our results with Cohan et al. (2018), who created the SMHD dataset. Because the authors have used the complete dataset and based on the assumption that the reported F1 score is computed on the positive class, we could not directly compare our results with the results published by the authors. However, using a randomly selected sample from the SMHD dataset for train, validation and test, we achieved a macro F1 score of 0.875 for predicting users with PTSD. In comparison with the binary classification results reported by Cohan et al. (2018), our proposed architecture has produced a significant improvement. Using a supervised FastText model, the authors have produced an F1 score of 0.576 for the binary classification task of predicting users diagnosed with PTSD. For the rest of the binary classification tasks to predict users with depression, ADHD, anxiety, bipolar, autism, OCD and Schizophrenia, the reported F1 scores by the authors are < 0.60. On the contrary, the lowest results we obtained when predicting users diagnosed with a mental disorder (i.e., when diagnosed with single or multiple mental disorders) is for ADHD with an F1 score of 0.688. The score was obtained for the urgent/not urgent task when used with the content from users diagnosed with a single mental disorder. The majority of the results (i.e., the F1 score from the eight disorders) for the classification tasks to identify users diagnosed with a mental disorder (i.e., primary diagnosis) is > 0.70.

5.7 Summary

Overall, it is clear that mental disorder detection significantly impacts suicide ideation detection when used in a multi-task learning environment with mixed parameter sharing. We could identify a substantial impact on suicide ideation detection when using PTSD mental disorder detection as one of the binary classification tasks. When taking the models that achieved a macro F1 score above 0.8 for suicide ideation detection, it could be identified that PTSD (0.865), bipolar (0.859) and schizophrenia (0.823) have shared more hidden features with suicidal thoughts. Similarly, for the mental illness detection, the disorders PTSD, bipolar and schizophrenia reported F1 scores 0.865, 0.846 and 0.835. Even though the rest of the mental disorders have also shared knowledge with suicide ideation, the tasks demonstrating effective bilateral knowledge sharing have formed well-generalized models. Similar behaviour was identified when predicting users with suicide ideation who require urgent attention. The disorders PTSD (0.862), anxiety (0.842), schizophrenia (0.837), autism (0.818) and bipolar (0.802) have shared more hidden features with suicide ideation compared to the remaining mental disorders. Similarly, for the mental illness detection task, PTSD (0.862), schizophrenia (0.845) and bipolar disorders (0.824) have gained more knowledge from the shared features with the suicide ideation detection task.

One of the key findings from our experiments is the presence of comorbidity of mental disorders and the impact it had on suicide ideation detection. Comorbidity is the presence of more than one medical condition in a patient at a given time. The presence of comorbid illnesses was identified among suicide victims with bipolar disorder (Simpson & Jamison, 1999) and individuals with severe depression (Handley et al., 2018). In general, comorbidities increase the risk of suicide irrespective of the

mental disorder (Bachmann, 2018). The significant impact of comorbid disorders was identified throughout the experiments.

Compared to single mental disorders' impact on suicide ideation and mental illness detection, comorbid disorders have further enhanced model performances when predicting users with suicide ideation or mental disorders. For example, when predicting users with and without suicide ideation, the disorders combined with suicide data that generated the best predictions were comorbid disorders where the primary diagnosis was PTSD followed by bipolar disorder. The impact of comorbidity was not detected only with suicide ideation detection tasks but also with mental illness detection where for both the flagged/not flagged and urgent/not urgent tasks, most of the best performances were obtained when using comorbidity of mental disorders. For our future work, we will conduct further research into identifying the particular disorders within the comorbidity of mental disorders that has more impact on enhancing suicide ideation detection.

Chapter 6

Cross-Platform Knowledge Transfer

The use of multi-task learning (MTL) is to share hidden features between the related tasks. We identified throughout our previous experiments that suicide ideation and mental disorder detection tasks do share hidden features. As a result, we managed to achieve well-generalized models (i.e., for the two tasks, flagged/not flagged and urgent/not urgent) that produced improved results compared to a majority class baseline and a strong baseline based on single-task learning. One of the extensively proven deep learning approaches to share knowledge between similar tasks is transfer learning. With transfer learning, one or more pre-trained layers on a similar task can be used with the task of interest so that certain low-level features can be shared between the two tasks (Goodfellow et al., 2016). Even with the proven success of applying transfer learning methods on data obtained from different social media platforms, it has not produced reliable results when using data annotated with proxy-based methods (Harrigian et al., 2020). This chapter aims to identify the adaptability of cross-platform knowledge sharing using multi-task learning instead of transfer learning. To accommodate our objective, we will be using the proposed architectures in sections 5.2.1, 5.2.2, and 5.2.3, which are based on multi-task and single-task

learning architectures. For all the experiments mentioned below, we used the UMD dataset and the CLPSych 2015 dataset as mentioned in chapter 3, section 3.1.3 (The University of Maryland Reddit Suicidality dataset) and section 3.1.2 (Twitter mental illness detection dataset). The experiments section below states in detail the different permutations of data samples that we have used for our experiments. Due to the limited number of data points provided with the CLPSych 2015 Twitter dataset, we could not test the impact individual mental disorders have on suicide ideation detection, but only the impact as a collection of mental disorders has on suicide ideation detection and also the suicide ideation detection task has on mental disorders as a collective. Similar to the experiments conducted in chapter 5, section 5.3, we will be using the proposed architecture with the auxiliary inputs to identify the impact EMPATH categories have on the model performances. The baseline architecture (i.e., figure 5.2.3) (i.e., using single-task learning) will be used to generate the baseline scores for mental illness detection.

6.1 Experiments

6.1.1 Task: Flagged/not Flagged

Similar to the experiments conducted using the UMD and SMHD datasets (Chapter 5), we grouped the users from the "Low", "Moderate" and "Severe" risk categories into the positive class and the users in the "No" risk category into the negative class. The number of instances between the positive and negative classes is balanced by adding 242 randomly selected UMD control users to the negative class. The finalized dataset to detect users with suicide ideation contained a total of 738 users equally divided among positive and negative classes. The test dataset was kept untouched,

which includes 125 users, where 80 users were categorized into the positive class while 45 users were grouped into the control class. To match the binary classes extracted from the UMD dataset, we randomly selected stratified samples from the CLPSych 2015 dataset for the mental illness detection task. Table 6.1 mentioned below lists the conducted experiments and the number of samples taken from each mental disorder.

Tasks	**Training**			**Testing**		
	Depress	**PTSD**	**Control**	**Depress**	**PTSD**	**Control**
suicide + more_ptsd	173	196	369	44	49	32
suicide + more_depress	257	112	369	65	28	32
suicide + ptsd_depress	210	159	369	53	40	32
suicide + ptsd_depress + embed	210	159	369	53	40	32
suicide + ptsd_depress + empath	210	159	369	53	40	32
baseline (ptsd + depress)	210	159	369	53	40	32

Table 6.1: The number of samples taken for each task to detect suicide ideation and mental illness using UMD and CLPSych 2015 data

The data is selected proportionately to the original CLPSych 2015 dataset. Each of the experiments as mentioned before and their differences are as follows:

- suicide + more_ptsd: Proportionally using more data from users diagnosed with PTSD for the mental illness detection task to be used in parallel with the suicide ideation detection task.
- suicide + more_depress: Proportionally using more data from users diagnosed with depression.
- suicide + ptsd_depress: Proportionally using data from users diagnosed with depression and PTSD.

- suicide + ptsd_depress + embed: Proportionally using data from users diagnosed with depression and PTSD combined with pre-trained word embeddings.
- suicide + ptsd_depress + empath: Proportionally using data from users diagnosed with depression and PTSD combined with EMPATH categories.
- baseline (ptsd + depress): Proportionally using data from users diagnosed with depression and PTSD to compute the baseline for the mental illness detection task.

To measure the impact word embeddings have on model performance, we used pre-trained fastText word embeddings with 300 dimensions. We made the embedding weights trainable so that the parameters of the embedding layer could get retrained according to the task. To further identify the possible impact exploratory analysis outcome of the EMPATH categories using the combined CLPSych 2015 and UMD data (Subsection 3.3.2) on the proposed multi-task learning architecture, we merged the following 10 EMPATH categories as auxiliary inputs to the model.
"health", "medical_emergency", "crime", "horror", "war", "sadness", "fear", "suffering", "aggression", "neglect"
We used the EMPATH categories only on the best-performing model to enhance the model performances. We did not conduct extensive experiments by using different combinations of the filtered categories or using different hyperparameters associated with the auxiliary inputs, as our main objective is to identify the impact mental disorders have on suicide ideation detection.

6.1.2 Task: Urgent/not Urgent

The main difference between the previous task (i.e., flagged/not flagged) and the current one is that we combine "No" and "Low" risk users from the UMD dataset to form the control class while the "Moderate" and the "Severe" risk categories are combined to form the positive class. To balance the positive and the negative classes, we randomly selected 142 users from the UMD control group and combined them with the negative class to generate the final dataset containing 319 users. Collectively 638 users were used for training and validation. The test dataset was kept untouched as before, where from 125 users, 80 users were categorized into the positive class while the remaining 45 users were grouped into the control class. To match the suicide ideation detection dataset, we randomly selected stratified samples from the CLPSych 2015 dataset in line with the experiments mentioned below in table 6.2.

Tasks	**Training**			**Testing**		
	Depress	**PTSD**	**Control**	**Depress**	**PTSD**	**Control**
suicide + more_ptsd	123	196	319	31	49	45
suicide + more_depress	257	62	319	65	15	45
suicide + ptsd_depress	181	138	319	45	35	45
suicide + more_ptsd + embed	123	196	319	31	49	45
suicide + more_ptsd + empath	123	196	319	31	49	45
baseline (more_ptsd + depress)	123	196	319	31	49	45

Table 6.2: The number of samples taken for each task to detect users with suicide ideation (i.e., who requires urgent attention) and mental illness using UMD and CLPSych 2015 data.

We conducted the same first three experiments as mentioned before in table 6.1. The only difference that we identified for the remaining set of experiments is, increasing the sample size of PTSD data rather than using stratified samples of users diagnosed with PTSD and depression. The remaining set of experiments are as follows:

- suicide + more_ptsd + embed: Proportionally using more data from users

diagnosed with PTSD and the remaining from those diagnosed with depression and combined with pre-trained word embeddings.

- suicide + more_ptsd + empath: Proportionally using more data from users diagnosed with PTSD and the remaining from those diagnosed with depression and combined with EMPATH categories.

- baseline (more_ptsd + depress): Proportionally using more data from users diagnosed with PTSD and the remaining from those diagnosed with depression to compute the baseline for the mental disorder detection task.

To identify the impact pre-trained embeddings have on model performances, we used pre-trained fastText word embeddings with 300 dimensions and made the embedding weights trainable. Similar to the flagged/not flagged task, we used the same EMPATH categories to discover the possibility of using auxiliary inputs to enhance the model performances. The EMPATH categories were only used with the best-performing model.

6.1.3 Creating the Vocabulary

To create the vocabulary, we used the same approach used when detecting suicide ideation and mental disorders using the UMD and SMHD datasets (Section 5.3.3). We created the vocabulary by merging the UMD and CLPSych 2015 training data. The maximum sequence length was calculated by taking the value five standard deviations away from the mean sequence length calculated using the merged training datasets. Because of the random selection of samples from the CLPSych dataset, the mean sequence length in the merged dataset is bound to change, and hence the

maximum sequence length will also change. Given the newly computed maximum sequence length, we normalized the sequence lengths that are longer than the new maximum sequence length by removing tokens from the "front", "end", or from "random" positions within the sequence. For the flagged/not flagged task, we identified that the best performances could be obtained by removing tokens from the "end" of the sequence, and for the urgent/not urgent task, it is from the "front" of the sequence. To generate the vocabulary, we took all the tokens available after merging the training datasets. Mentioned in table 6.3 are the computed maximum sequence lengths and the vocabulary size for each experiment conducted under the flagged/not flagged task.

Tasks	**New maximum sequence length**	**Vocabulary size**
suicide + more_ptsd	45,828	211,443
suicide + more_depress	45,461	216,726
suicide + ptsd_depress, suicide + ptsd_depress + embed, suicide + ptsd_depress + empath, baseline (ptsd + depress)	45,593	216,554

Table 6.3: Selected vocabulary size and the newly calculated maximum sequence length according to the given task

For the urgent/not urgent task, we selected the most frequent 200,000 words identified from the merged training dataset. Our experiments discovered that using a vocabulary with 200,000 most frequently used tokens produces better results than a vocabulary with more or fewer tokens. The maximum sequence lengths to be used varied from task to task where for the tasks with more PTSD users it was 46,483 tokens, for the task with more depressed users, it was 46,030 tokens, and for the task that contained a stratified sample of users (i.e., from both depressed and PTSD users) it was 45,836.

6.1.4 Baseline

We used the same baseline scores for suicide ideation detection, identified in chapter 5 subsections 5.4.1 (i.e., for the task flagged/not flagged) and 5.4.2 (i.e., for the task urgent/not urgent). Suicide ideation detection baseline F1 score for flagged/not flagged task is 0.654, and for the urgent/not urgent task, it is 0.616.

To calculate the baselines for mental illness detection within the flagged/not flagged task, we used a stratified sample of users diagnosed with PTSD and depression. For the urgent/not urgent task, we selected more instances of the PTSD class during the training phase because, from several preliminary experiments, we identified that using more users diagnosed with PTSD compared to the number of users diagnosed with depression produced better results. For example, when calculating the baseline for the task urgent/not urgent, we used 196 users diagnosed with PTSD and 123 users diagnosed with depression. A total of 319 users were used as the group where users are diagnosed with mental illnesses (i.e., users with PTSD + depression). We selected 49 PTSD users and 31 depressed users to be used jointly as the positive class for inference. The control class contained 45 neurotypical users.

6.1.5 Model Training

Similar to the experiments conducted with the UMD and SMHD datasets to predict suicide ideation and mental disorders, we used five stratified shuffle splits where 80% of the data is used for training and 20% for validation. We made sure that the tasks to be trained are not misaligned so that a user with suicide ideation is in line with a user diagnosed with a mental illness, which could be either PTSD or depression. If not aligned, the tasks will not be able to share common features among users with suicide

ideation and mental disorders and also, it will be difficult for the model to distinguish neurotypicals from the ones with suicidal thoughts and mental illnesses. The same proposed architecture (Subsection 5.2.1) used when predicting users with suicide ideation and mental disorders (i.e., using the UMD and SMHD datasets) was used with minor changes to the hyperparameters. We used a learning rate of 0.001 with Adam optimizer (Kingma & Ba, 2015). After experimenting with several activation functions, we discovered LeakyReLU (Maas et al., 2013) to be more reliable when used as the activation function with an alpha(α) value of 0.2 on the outputs generated by the convolution and dense layers. To reduce model overfitting, we used dropout (Srivastava et al., 2014) with a probability of 0.5 and $L1$ and $L2$ regularization with a regularization factor (10^{-5}) to penalize convolution and fully connected layers for having larger weights. We used a custom loss function, which summed categorical cross-entropy loss and mean squared error, and was minimized during the training phase. Given the two different datasets, we experimented with several batch sizes to train the model and identified a mini-batch of size 8 to produce better results than larger batch sizes. We trained the model for 10 epochs with early stopping (i.e., only if the validation loss did not improve for 3 epochs). In addition, we reduced the learning rate by a factor of 0.1 if the validation loss did not improve for 2 consecutive epochs. The minimum learning rate was initialized to be 10^{-8}. The model with the lowest validation loss was returned to be used at inference.

6.1.6 Inference

The same crowdsourced annotated UMD test dataset used with suicide ideation and mental illness detection (i.e., when using the UMD and SMHD datasets) is used at

inference. We did not conduct further testing using the UMD expert annotated test data due to the limited number of instances available with the CLPSych 2015 dataset. To match the imbalanced UMD test dataset, we randomly selected an equal number of instances from the Twitter data to be used as the test dataset (tables 6.1 and 6.2). The performance is evaluated on macro precision, recall and F1 score, and in addition, we also use the macro averaged ROC AUC score and prediction accuracy to understand the level of generalization the model has achieved. Given the stochastic nature of the deep learning algorithms, it is inevitable to have variance in the predicted results. Due to this reason and to derive a more reliable prediction, we used the model averaging ensemble approach (Brownlee, 2018) to make the final prediction on the test data. Using model averaging creates the opportunity to generalize the prediction outcome, where using different models trained on the same data might not make the same errors on the test data (Goodfellow et al., 2016).

6.2 Results

We conducted several experiments using distinct combinations of stratified samples to identify the impact PTSD and depression have on suicide ideation and mental illness detection. For each of the tasks that are either flagged/not flagged or urgent/not urgent, we selected a different number of instances from each mental disorder (i.e., from either PTSD or depression) to identify the level of impact each has on suicide ideation and mental illness detection. In addition, we wanted to identify if the impact of mental disorders recognized when using the UMD and SMHD datasets from the same social media platform (i.e., results from section 5.4) in chapter 5 can also be identified when using data from different social media platforms. For example, we

discovered that users whom self-declared PTSD share more hidden features with users having suicidal thoughts than other mental disorders such as depression.

We used pre-trained embeddings and EMPATH categories with the best-performing model to see if the trained model could be further generalized given the test data. We used the same stratified sample that produced the best macro F1 score with multi-task learning to compute a strong baseline. We highlighted the best results, emphasized the best F1 score, and highlighted the results if improved with either the EMAPTH categories or the pre-trained word embeddings.

The results tables mentioned in the following sections consist of the same metrics as mentioned in tables 5.3 and 5.4 in chapter 5.

6.2.1 Task: Flagged/not Flagged

Table 6.4 states the results obtained for each of the experiments conducted to predict users with suicide ideation and mental disorders (i.e., the generic category of having a mental illness or not). Each task is given with different data combinations to identify the impact different mental disorders (i.e., PTSD or depression) have on suicide ideation detection and suicide ideation detection task has on mental illness detection. The two baselines are to identify if a user has a mental disorder or not and if a user is having suicidal thoughts or not.

Multi-task	**Ps**	**Pm**	**Rs**	**Rm**	**F1s**	**F1m**	**ACCs(%)**	**ACCm(%)**	**AUCs**	**AUCm**
suicide + more_ptsd	0.857	0.857	0.842	0.842	0.849	0.849	88.80	88.80	0.956	0.955
suicide + more_depress	0.793	0.783	0.851	0.835	0.810	0.799	84.00	83.20	0.928	0.928
suicide + ptsd_depress	0.861	0.861	0.868	0.868	0.864	0.864	89.60	89.60	0.960	0.960
suicide + ptsd_depress + embed	0.846	0.836	0.873	0.868	0.858	0.849	88.80	88.00	0.946	0.946
suicide + ptsd_depress + empath	0.869	0.869	0.884	0.884	**0.876**	**0.876**	90.40	90.40	0.960	0.959
baseline (ptsd + depress)	-	0.806	-	0.877	-	0.824	-	84.80	-	0.951
Suicide baseline	0.647	-	0.673	-	0.654	-	71.20	-	0.726	-

Table 6.4: Results obtained for the task flagged/not flagged with different sample sizes of users diagnosed with PTSD or depression

6.2.2 Task: Urgent/not Urgent

Table 6.5 lists the experiments conducted to identify if a user has suicidal thoughts (i.e., requiring urgent attention) and mental illness. Each task in the multi-task column is submitted with different combinations of inputs to identify the impact different mental disorders have on suicide ideation and the suicide ideation detection task has on mental illness detection. In addition, two baseline predictions are included to identify users with suicide ideation who requires urgent attention and users with mental disorders, which could be either PTSD or depression.

Multi-task	Ps	Pm	Rs	Rm	F1s	F1m	ACCs(%)	ACCm(%)	AUCs	AUCm
suicide + more_ptsd	0.878	0.878	0.907	0.907	0.884	0.884	88.80	88.80	0.959	0.959
suicide + more_depress	0.814	0.814	0.841	0.841	0.812	0.812	81.60	81.60	0.925	0.925
suicide + ptsd_depress	0.866	0.866	0.881	0.881	0.872	0.872	88.00	88.00	0.920	0.920
suicide + more_ptsd + embed	0.866	0.866	0.886	0.886	0.873	0.873	88.00	88.00	0.924	0.925
suicide + more_ptsd + empath	0.883	0.891	0.899	0.910	**0.889**	**0.898**	89.60	90.40	0.946	0.945
baseline (more_ptsd + depress)	-	0.840	-	0.868	-	0.829	-	83.20	-	0.965
Suicide baseline	0.616	-	0.625	-	0.616	-	63.20	-	0.680	-

Table 6.5: Results obtained for the task urgent/not urgent with different sample sizes of users diagnosed with PTSD or depression

6.3 Discussion

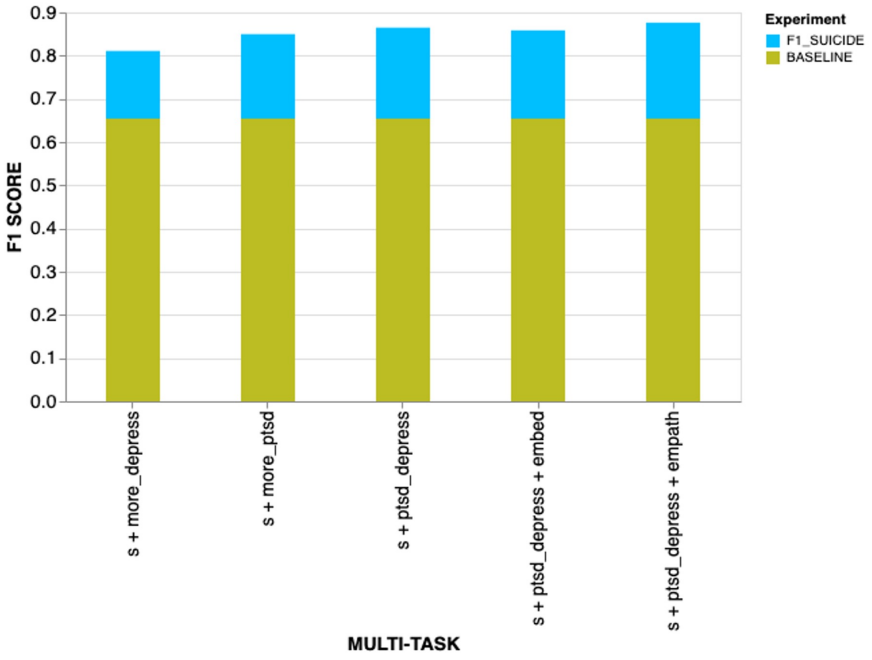
To analyze the overall outcome of our experiments in detecting suicide ideation and mental disorders using cross-platform data, we have included Figures 6.1, 6.2, 6.3, and 6.4 that demonstrate the combination of tasks in the x-axis with their associated macro F1 scores in the y-axis. The outcome of each experiment is represented in the layered bar chart, where the labels mentioned in the chart legend indicate the F1 score for the related experiment:

- F1_SUICIDE: F1 score for suicide ideation detection task (i.e., for flagged/not flagged or urgent/not urgent).

- BASELINE: Baseline F1 score for suicide ideation / mental illness detection (i.e., for flagged/not flagged or urgent/not urgent).
- F1_MENTAL: F1 score for mental illness detection task (i.e., for flagged/not flagged or urgent/not urgent).

6.3.1 Task: Flagged /not Flagged

Analyzing table 6.4 and figure 6.1, we could see that using mental illness data has significantly improved prediction results of users with suicide ideation.



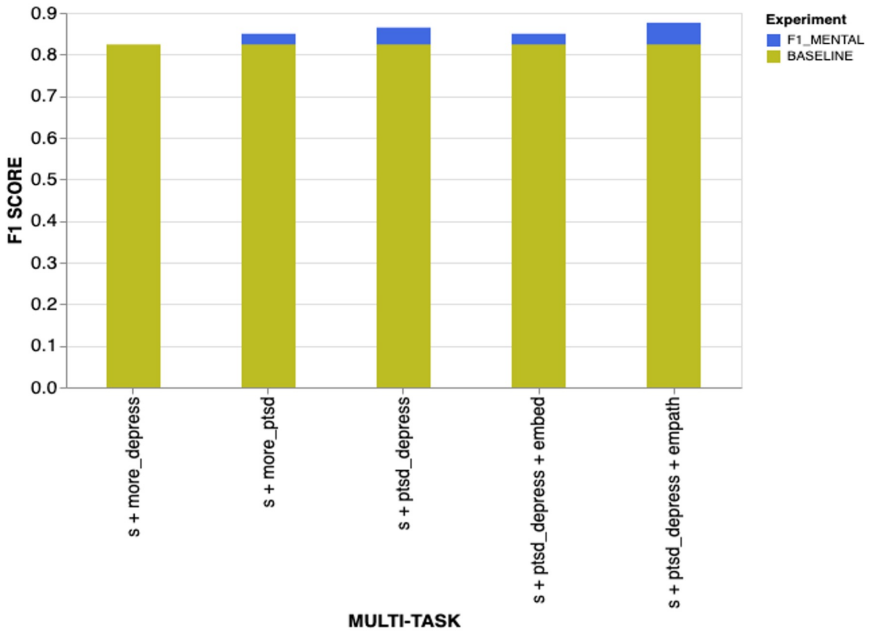
	suicide + more_depress	suicide + more_ptsd	suicide + ptsd_depress	suicide + ptsd_depress + embed	suicide + ptsd_depress + empath
F1_SUICIDE	0.81	0.849	0.864	0.858	0.876
BASELINE	0.654	0.654	0.654	0.654	0.654

Figure 6.1: Overall results for the suicide ideation detection task with users diagnosed with mental disorders(i.e., either PTSD or depression). Also, includes the suicide ideation detection baseline.

Each bar in the overlayed bar chart is associated with the table columns mentioned in figure 6.1. For example, the x-axis label "s + more_depress" represents the multi-task that used suicide ideation and mental disorder data where most users with a mental disorder are diagnosed with depression. Similarly, the bar chart label prefix "s +" denotes the inclusion of data that represents users with suicide ideation. When detecting users with suicide ideation, all the experiments we have conducted have surpassed the baseline F1 score of 0.654. The best F1 score (0.876) is reported for the experiment that used a stratified sample of users diagnosed with PTSD or depression with the EMPATH categories as auxiliary inputs. The auxiliary inputs have improved the F1 score by around 1% compared to when not using the auxiliary inputs. The best F1 score without the auxiliary inputs is 0.864. In addition, we could identify that using more data from users diagnosed with PTSD has produced a better F1 score (0.849) than when using more data from users diagnosed with depression (0.810). However, based on the best results obtained for suicide ideation detection, it could be identified that more common features are being discovered when using data from multiple mental disorders rather than taking more data from a single disorder.

According to figure 6.2, it is clear that suicide ideation detection tasks have also had a noticeable influence on the mental illness detection task where when using data from users diagnosed with either PTSD or depression have improved the mental illness detection F1 score from a strong baseline of 0.824 to 0.864 (i.e., s + ptsd_depress).

Similar to figure 5.7 results when using UMD and SMHD data, we could identify that using more data from users diagnosed with depression has not improved upon the mental illness detection baseline. Similar to figure 6.1, the best F1 score for the mental illness detection task is obtained when combining data from users diagnosed with



	suicide + more_depress	suicide + more_ptsd	suicide + ptsd_depress	suicide + ptsd_depress + embed	suicide + ptsd_depress + empath
F1_MENTAL	0.799	**0.849**	**0.864**	**0.849**	**0.876**
BASELINE	**0.824**	0.824	0.824	0.824	0.824

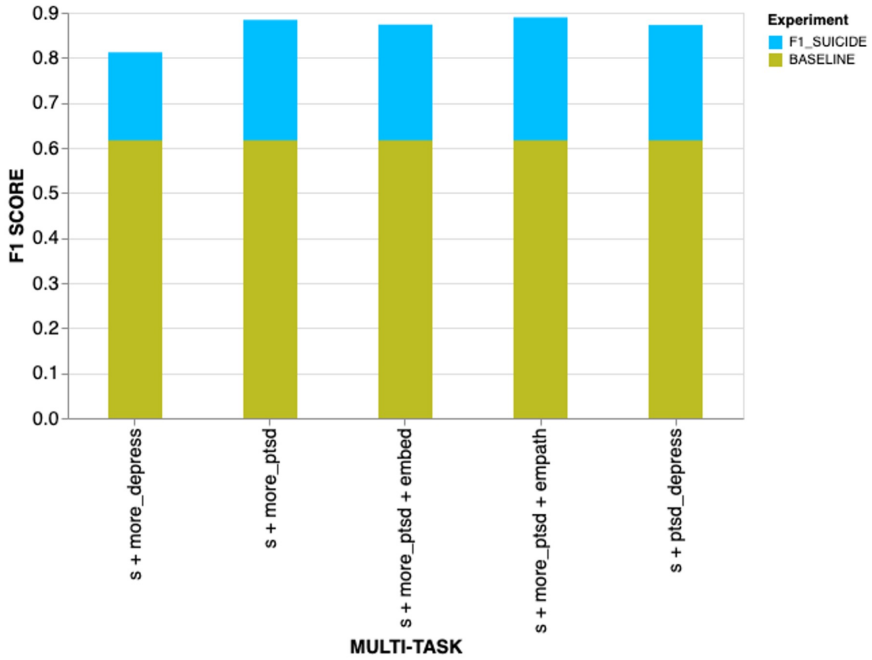
Figure 6.2: Overall results for the mental illness detection task with data from users diagnosed with either PTSD or depression. It also includes the mental illness detection baseline.

PTSD or depression with the auxiliary inputs (i.e., 0.876). Considering the UMD and CLPSych 2015 data, we could see that, even though data from users diagnosed with depression improved the performance of the suicide ideation detection task over its baseline predictions, the features being shared between users diagnosed with depression and users having suicidal thoughts are fewer in comparison to when using more users diagnosed with PTSD.

6.3.2 Task: Urgent/not Urgent

According to table 6.5 and figure 6.3, a considerable improvement over the suicide ideation detection baseline (i.e., 0.616) can be identified with all the experiments conducted using the CLPSych 2015 data. The highest F1 score of 0.889 is obtained using the suicide ideation data combined with more data from users diagnosed with PTSD and the EMPATH categories as auxiliary inputs. Using the EMPATH categories has improved the macro F1 score by 1% compared to not using the categories, which produced an F1 score of 0.884. Overall, all the experiments have achieved a macro F1 score above 0.80, and a considerable improvement can be identified when using more data from users diagnosed with PTSD. A lower impact can be identified when using more data from users diagnosed with depression, with the lowest F1 score (0.812) for predicting users with suicide ideation. Overall, when predicting users with suicide ideation who require urgent attention, the features contributed by those diagnosed with PTSD have significantly improved the models' predictability. The impact contributed by these features on suicide ideation detection has increased from predicting users with suicide ideation (i.e., the flagged/not flagged task) to detecting users who require urgent attention (i.e., the urgent/not urgent task).

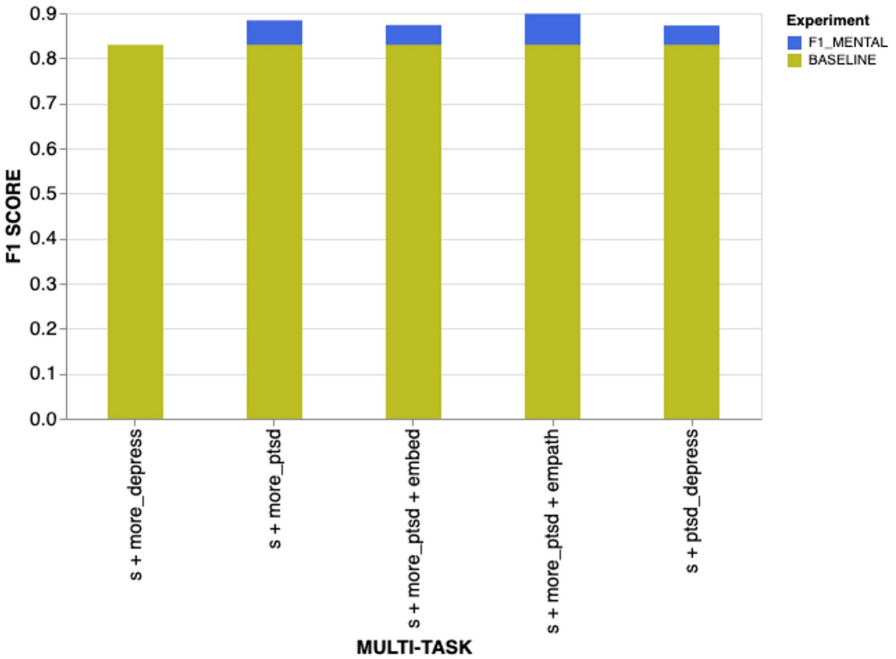
According to figure 6.4, we could identify the impact suicide ideation detection task has on mental illness detection where except for one experiment, the reported performances of the rest of the experiments have improved from their respective baseline. The highest F1 score is reported when using more data from users diagnosed with PTSD and merged with the EMPATH categories. Without using the auxiliary inputs, the model has reported a macro F1 score of 0.884, which is about 1.4% less than the highest F1 score of 0.898. Similar to the previous outcomes depicted in



	suicide + more_depress	suicide + more_ptsd	suicide + more_ptsd+ embed	suicide + more_ptsd + empath	suicide + ptsd_depress
F1_SUICIDE	0.812	0.884	0.873	0.889	0.872
BASELINE	0.616	0.616	0.616	0.616	0.616

Figure 6.3: Overall results for the suicide ideation detection (i.e., for users requiring urgent attention) with data from users diagnosed with PTSD or depression. It also includes the suicide ideation detection baseline.

figures 5.7, 5.14 and 6.2, the impact suicide ideation detection task has on predicting users with depression is considerably low. Even though we have computed a strong baseline using more data from users diagnosed with PTSD, the impact of having more data from users diagnosed with depression on the overall model performance is comparatively low. With both tasks reporting an F1 score of 0.812, it could be derived that the two tasks (suicide ideation and mental illness detection) share the least number of hidden features when using more data from users diagnosed with depression.



	suicide + more_depress	suicide + more_ptsd	suicide + more_ptsd+ embed	suicide + more_ptsd + empath	suicide + ptsd_depress
F1_MENTAL	0.812	**0.884**	**0.873**	**0.898**	**0.872**
BASELINE	**0.829**	0.829	0.829	0.829	0.829

Figure 6.4: Overall results for the mental illness detection task with data from users diagnosed with either PTSD or depression. It also include the mental illness detection baseline.

6.4 Comparison to Related Work

During our experiments on cross-platform knowledge transferring, where we used the CLPSych 2015 data for the mental illness detection task, we identified that in comparison to when using the SMHD data, the performances of the suicide ideation detection task has improved. For the flagged/not flagged task, the improvement was when using a stratified sample of users from the original dataset diagnosed with PTSD and depression. However, the best results were obtained for the urgent/not urgent task when extracting a sample that contained more users diagnosed with PTSD than

depression. Table 6.6 demonstrates our best results compared to the best results reported by the CLPSych 2019 task participants.

Submissions	F1(flagged/not flagged)	F1(urgent/not urgent)
Matero et al. (2019)	0.821	0.816
CAMH	**0.91**	0.812
Iserman et al. (2019)	0.848	0.775
Mohammadi et al. (2019)	0.843	0.718
MTL with mixed parameter sharing(using Twitter data)	0.876	**0.889**

Table 6.6: Related work comparison for suicide ideation detection(using Twitter data)

Similar to table 5.17, we have obtained better results over the best results reported by the CLPSych 2019 task participants for the urgent/not urgent task. Our flagged/not flagged task results are ranked second with a macro F1 score of 0.876. Compared to the results obtained using the SMHD dataset (i.e., according to table 5.17), the macro F1 score for the flagged/not flagged task has improved from 0.875 to 0.876, and for the urgent/not urgent task, the improvement is from 0.862 to 0.889.

To get an overall understanding of the level of performance our proposed model has achieved when using the CLPSych 2015 dataset, we can compare our results with the CLPSych 2015 shared task results. According to the AUC scores: 0.86 (depression vs. control), 0.84 (depression vs. PTSD) and 0.89 (PTSD vs. control) reported by Resnik, Armstrong, Claudino, and Nguyen (2015), we have obtained better results when detecting users with mental disorders (i.e., PTSD or depression). For the tasks flagged/not flagged and urgent/not urgent, our proposed model has generated AUC scores 0.959 and 0.945, respectively. Unlike in chapter 4, where we only used validation data to measure how well the trained model has generalized, in this chapter, we randomly selected a test dataset to measure the performance of our trained model on unseen data. Due to the inadequate number of instances and architectural requirements, we could not conduct our research according to CLPSych 2015 shared

task requirements, and if so, our research outcome could have been compared directly with the CLPSych 2015 shared task participants. Instead, we merged the users with PTSD and depression to one category labelled as having mental disorders or not to identify the general impact mental illnesses have on suicide ideation and the impact suicide ideation detection has on mental illness detection. Similar to the results obtained using the SMHD dataset, we could identify that PTSD has a significant impact on suicide ideation detection, and also, the suicide ideation detection task shares more hidden features with users diagnosed with PTSD than with other mental disorders.

6.5 Summary

According to the research outcome, it can be identified that cross-platform data can be used successfully in a multi-task learning environment. The proposed multi-task learning architecture with both hard and soft parameter sharing has managed to detect shared features between the tasks and segregate the task-specific features. The outcomes resemble similarities discovered when using the UMD and SMHD datasets from the same social media platform. Overall, we identified that users diagnosed with depression or PTSD have collectively shared features with suicide ideation that managed to distinguish neurotypicals from the ones having suicidal thoughts. Even though comorbidities are identified in an individual, the overall capability of collectively using more than one mental disorder to share hidden features with suicide ideation validates the impact comorbidity of mental disorders has on improving suicide ideation predictions.

Our results also indicate that more similarities can be identified between the users

with moderate to severe suicide risk and a group of users where the majority are diagnosed with PTSD. When increasing the number of users diagnosed with depression, the overall performance of the model decreased. The decline can be subject to several reasons where, given the dataset, users identified as having depression might not possess any risk of suicide. According to Hawton et al. (2013), several risks factors were identified that could increase the chances of suicide, such as severe depression, hopelessness, previously attempted suicide and comorbid disorders. Hypothetically, it could be argued that, unless the users diagnosed with depression have reflected the associated risks of suicide in their posts, the features shared between users with suicide ideation and depression will be less informative. Concerning the use of auxiliary inputs, we identified that the performance of the proposed model could be enhanced by using inputs such as EMPATH categories discovered through exploratory analysis.

Chapter 7

Conclusion and Future Work

7.1 Applications

Given the sensitive nature of the data and especially with the ethical and privacy concerns, it is crucial to understand the factors that must be considered when introducing specific solutions into clinical environments. For example, detecting an individual with suicide risk is insufficient unless linked with an intervention mechanism (Linthicum et al., 2019). Over the years, only a few applications were implemented successfully to predict users with suicide ideation and mental disorders as an initial part of the intervention mechanism. Two of the applications that failed to protect users' privacy were Samaritans Radar (Samaritans, 2015) and Koko (Jaroszewski et al., 2019). The Samaritans Radar was a Twitter plugin that allowed Twitter users to monitor each others' posts. It matched keywords and phrases and sent out emails to the users who had registered to monitor that account. The app was criticized for breaching user privacy and was suspended after nine days from launch. Similarly, Koko and Foveocare are applications that individuals can use to interact with each other to discuss their problems but were not supervised by a health care professional except for

being monitored by moderators (Skenderi, 2016). Even though it could be beneficial to discuss your current situation with a community of users who have faced similar situations, users in distressful circumstances need to receive professional guidance to reduce further mental and physical strains. Our hope for building a reliable platform for identifying individuals with suicide ideation and mental disorders is to discover opportunities to deploy our solution to initiate a mechanism for people in need of mental health support to obtain the professional care they require.

As a successful application that used natural language processing and machine learning to identify individuals with mental disorders and suicide ideation, Milne et al. (2019) introduced a machine learning classifier to prioritize forum posts based on their severity. The posts were prioritized based on four categories, green, amber, red and crisis, where the site moderators will attend to the users based on the severity level. After implementing the proposed solution, the response delays for the messages labelled as a crisis, red, amber and green were reduced by 80%, 80%, 77% and 12%.

When considering deployment opportunities for our proposed architecture, it is essential to consider the privacy concerns and the ethical implications that such an application could introduce into the suicide ideation and mental illness detection domain. Similar to Milne et al. (2019), we could introduce our proposed model to a platform where consenting users can visit for mental health support and will get prioritized based on the level of risk associated with their posted content. As mentioned before, these platforms must be regularized by trained moderators who will handle users with less critical concerns, while qualified mental health professionals will handle severe incidents. In addition to these support forums, we will investigate the avenues of using our proposed architecture in other platforms (e.g., subreddits not

related to mental health) under strict ethical and privacy guidelines.

7.2 Conclusion

This research highlights the importance of early detecting users with mental disorders and suicide ideation. We proposed several multi-task learning architectures to identify at-risk users and managed to produce state-of-the-art results in detecting users with mental disorders and suicide ideation. It was identified that many Canadians could not receive the necessary support they need for their mental health. Without the required mental health support, one's mental and physical health could deteriorate. Two of the main reasons many have not received mental health support were that they could not afford it or did not know where to get the necessary support. When analyzing the social media usage in Canada and its demographics, we identified that social media platforms could be used as a powerful tool to provide the necessary mental health care if ethical principles are respected. To provide the necessary support, first, it is essential to detect the individuals who require help. Using our proposed multi-task learning architecture, we proved that users with suicide ideation or mental disorders could be identified with low false positive and false negative rates.

In the process of detecting users with suicide ideation and mental disorders, we used four datasets to conduct three sets of experiments, as demonstrated in chapters 4, 5 and 6. In chapter 4, we demonstrated that we could successfully use emotion categories as auxiliary inputs to enhance the performances of the mental illness detection task. Multi-task learning with hard parameter sharing was used to identify three categories of users: neurotypicals, PTSD and depression. The MTL architecture consisted of shared layers to automatically discover the shared features between the tasks and dedicated

layers to identify features unique to each task. The predicted emotion categories were used as auxiliary inputs when detecting neurotypicals and users diagnosed with depression, while age and gender were used to differentiate users diagnosed with PTSD from the rest of the users. The results were compared against a multi-class classification baseline that used an SVM algorithm. The reliability of our proposed architecture was further tested by comparing it against different architectures that used: pre-trained word embeddings, RNN based neural networks and multi-channel CNN without multi-tasking. When compared against the state-of-the-art results, our proposed architecture has produced better results, but direct comparisons could not be made since we did not have access to the test dataset provided to the CLPSych 2015 shared task participants.

With the success of using multi-task learning to identify mental disorders, the next objective of our research was early detecting users with suicidal thoughts given the Reddit social media data. Rather than implementing single-task learning to identify users with suicidal thoughts, we broadened our experiments by introducing mental illness detection to formulate a multi-task learning environment. The reason for including mental illness detection as one of the tasks in a multi-task learning environment was the thorough research conducted by public health researchers to identify the relationship between suicide risk and mental disorders. Our research used content published by Reddit social media users from the subreddit SuicideWatch and content from Reddit users who have self-declared one or more mental disorders. The input for the mental illness detection task was taken from the SMHD dataset with nine mental disorders, where we used only eight because the required number of users who self-declared eating disorders was not available. After experimenting

with different multi-task learning architectures that used hard parameter sharing, soft parameter sharing, and mixed parameter sharing, we identified that using hard and soft parameter sharing is more effective in discovering hidden shared features between users who self-declared mental disorders and having suicidal thoughts. In addition, we tested our best performing model with auxiliary inputs generated using the EMPATH categories and discovered that the overall performances could be improved, especially when predicting users with suicide ideation. On the contrary, the exact impact could not be identified when predicting users with suicide ideation who require urgent attention. The auxiliary inputs to be used were discovered during the exploratory analysis stage, and further research can be conducted to identify the benefits of using more categories or a combination of categories.

We used several pre-trained word embeddings throughout our research, but the best performances were obtained using randomly initialized embeddings. With extensive experiments, we discovered that when trying to identify users with suicide ideation, users diagnosed with comorbidity of disorders were sharing more features with users at risk of suicide than users diagnosed with a single mental disorder. Out of the eight tested disorders, users who self-declared PTSD, bipolar or schizophrenia as the primary diagnosis shared more features with users having suicidal thoughts (i.e., with a macro F1 score > 0.80). The same impact was identified when predicting users with mental disorders. Among the previously mentioned primary diagnoses, users whom self-declared schizophrenia have shared more features with suicide ideation only when diagnosed with a single mental disorder (i.e., only schizophrenia) than comorbidity of disorders. This unique outcome was only recognized from the content published by users who self-declared schizophrenia, and the impact can be identified from both

mental illness and suicide ideation detection tasks.

A similar outcome was recognized when predicting users with suicide ideation who require urgent attention, where users whom self-declared only PTSD shared more features with users having suicidal thoughts. The impact of comorbidity can be identified from the other mental disorders except from the users diagnosed with PTSD and depression. Using content from users diagnosed only with PTSD has produced an F1 score of 0.862 for suicide ideation and mental illness detection tasks. Even though users whom self-declared bipolar disorder have shared hidden features with users at-risk of suicide who requires urgent attention, more features were shared by users who self-declared anxiety, schizophrenia, and autism.

The models we trained using the UMD and SMHD datasets to predict users with suicide ideation and mental disorders were further tested with the UMD expert annotated dataset to identify how well our trained models have generalized. Through these tests, we identified that the trained models have generalized well on predicting users with suicide ideation and mental disorders with an F1 score > 0.84.

The final stage of our research was to identify the potentials of transferring knowledge between different data platforms (i.e., between Reddit and Twitter) with different distributions to predict users with suicide ideation and mental disorders. Due to the inadequate number of instances (i.e., from the CLPSych 2015 dataset), we could not conduct experiments using content from users diagnosed with a single mental disorder, and instead, we randomly sampled data from users diagnosed with PTSD and depression. Each sample consisted of either more or fewer users diagnosed with PTSD or depression and also with the same proportion of users diagnosed with either PTSD or depression as given in the original dataset. Similar to the experiments

conducted using the SMHD dataset, we identified that users with suicide ideation do share features with users diagnosed with either PTSD or depression. When predicting users with suicide ideation, we identified that using a sample consisting of an equal proportion of users as in the original dataset generated better performance than other combinations of samples. However, increasing the numbers of users diagnosed with depression in the sample reduced the model performance significantly (i.e., about 8% in F1 score), while increasing the number of users with PTSD reduced the model's performance by about 3%. Like our previous experiments, we could identify that users with different mental disorders share more features with suicide ideation than having more data from users diagnosed with a single mental disorder. Even though it could not be directly compared, the impact resembles the nature of being diagnosed with comorbidity of disorders. Being diagnosed with more than one mental disorder strongly correlates with suicide ideation compared to being diagnosed with a single mental disorder. However, when identifying users with suicide ideation who require urgent attention, the exact behaviour discovered when using the SMHD dataset was identified. Users diagnosed with PTSD were sharing more features with users at-risk of suicide who require urgent attention. Selecting a sample with more users who self-declared depression did not improve the model performance, and only when increasing the number of users diagnosed with PTSD improved the overall model performances. Using several selected EMPATH categories as auxiliary inputs also enhanced the prediction capabilities of the model where the maximum impact was identified when predicting users with mental disorders where the macro F1 score increased by 1% compared to when not using the auxiliary inputs.

Summarizing the overall results from chapters 5 and 6, we could see that out

of different mental disorders, PTSD has shared more features with users at risk of suicide, followed by schizophrenia, bipolar, anxiety, and autism. Irrespective of the data platform, suicide ideation and mental illness detection tasks have shared knowledge that has enhanced the early detection of users with mental disorders or suicidal thoughts. The impact mental disorders had on suicide ideation detection varied between different mental disorders. Overall, the hidden features from these disorders considerably improved suicide ideation prediction over its baseline, and produced state-of-the-art results when predicting users with suicide ideation who require urgent attention. The results aligned with public health research (Bachmann, 2018; Handley et al., 2018; Brådvik, 2018) that highlighted the significant impact mental disorders have on suicide ideation. We could not identify the same level of impact from the suicide ideation detection task on mental illness detection, where only a few mental disorders reported considerable improvement over their respective baselines.

Given the research outcome from chapters 5, 6 and 4, we can determine the answers to our research questions stated in section 1.3. To answer the first question, we could identify that multi-task learning with automatically-calculated auxiliary features can be used effectively to classify users with mental disorders. Specifically, emotion categories embedded within the posted messages have managed to distinguish users with mental disorders from neurotypicals. For question two, our experiments have provided sufficient evidence to demonstrate the effectiveness of multi-task learning in identifying mental disorders' impact (i.e., either single or multiple mental disorders) on suicide ideation detection and the impact suicide ideation has on mental illness detection, especially when the datasets are from two distributions. Finally, we

extended our research into multiple social media platforms, where we demonstrated the significant impact multi-task learning has on different but related tasks of suicide ideation and mental illness detection. We demonstrated sufficient evidence to prove that irrespective of the social media platform, both tasks have a noticeable impact on one another where more knowledge is transferred from mental illness detection to suicide ideation detection, therefore answering our third research question.

7.3 Limitations and Challenges

The main limitation faced by our research is not having access to enough data to use when training the model. Lack of training data can be identified throughout the training phase where the trained models for emotion, mental illness and suicide ideation detection were overfitting the training data. For example, when predicting the emotion categories to be included as auxiliary inputs, the model was trained on a limited number of data points, and as a result, we could not obtain a well-generalized model so that the overall prediction accuracies could be enhanced. Even though we managed to improve the mental disorder predictions (i.e., PTSD and depression), we could have obtained further improvements with better emotion predictions. Similarly, having more data to train the suicide ideation detection task could have also improved the overall predictions on the test data and also having more test data could have reduced the model variance. However, we obtained a well-generalized model using limited instances and different strategies to overcome model overfitting.

In our experiments, we could not use state-of-the-art transformer-based architectures due to the inability of the models to process longer sequences. For our research, we concatenated all the posts published by each user, and as a result, the average

sequence lengths per user for all the datasets were above the maximum sequence length expected by most of the transformer-based architectures. Even though the posts are concatenated according to their timestamps, we could not identify the exact locations within the sequence indicating the signs of mental disorders or suicidal thoughts. Due to this reason, truncating the longer sequences to accommodate the model requirements could significantly increase the number of false positives and false negatives. Even though specific models are being introduced to overcome the sequence length limitations, they are either not sufficient to accommodate the required sequence length for our experiments or even if it does manage to accommodate a longer sequence, it requires more computational resources to train the model (see section 2.1.2 for more details). Even though we did not use transformer-based architectures, we managed to obtain better results than research conducted using architectures such as BERT (e.g., Matero et al. (2019)).

7.4 Future Work

Using multi-task learning to predict users with mental disorders and suicide ideation has created further research opportunities. Given the tasks of predicting users with different mental disorders and suicide ideation, we can introduce more mental disorders into the mental illness detection task to identify their impact on suicide ideation detection, as well as the impact suicide ideation detection task has on newly introduced mental disorders. Apart from the hyperparameters we have fine-tuned to obtain optimal model performance, we can explore many other hyperparameters that can be optimized to enhance overall performance. In addition, we will conduct more research into the use of current state-of-the-art transformer-based architectures by exploring the methods

that can be used to overcome sequence length limitations.

Because our research did not focus on predicting the risk categories of users (i.e., "no", "low", "moderate" and "severe") due to our research objective and architectural requirements, we will conduct further research into predicting these risk categories by measuring the impact different mental disorders have on each category.

We will further investigate the UMD dataset by structuring it in a way that can be used to analyze the temporal change in the user-generated data. As people's mental health conditions could change over time, it is vital to assess the mental state of an individual (in our case, using text data) over a certain period rather than focusing on a limited time frame. The importance of temporal analysis was highlighted by Macavaney et al. (2018), where the annotations are based on diagnosis recency (i.e., when the users are diagnosed) and condition state (i.e. if the diagnosis is still current). Even though the before-mentioned temporal analysis considers the diagnosis of a user, it does not consider that individual's mental state within successive time frames. A user-based temporal analysis will be conducted in our future work by considering a series of posts indexed over time. In deriving the outcome (i.e., the risk of suicide), we will also detect if the user is diagnosed with any mental illnesses (i.e., within the given time frame) identified using a multi-task learning model trained on the SMHD data.

References

Abboute, A., Boudjeriou, Y., Entringer, G., Azé, J., Bringay, S., & Poncelet, P. (2014). Mining Twitter for suicide prevention. *Natural Language Processing and Information Systems*, 250–253.

Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual String Embeddings for Sequence Labeling. In *Coling 2018, 27th international conference on computational linguistics* (pp. 1638–1649).

Ambalavanan, A. K., Jagtap, P. D., Adhya, S., & Devarakonda, M. (2019). Using Contextual Representations for Suicide Risk Assessment from Internet Forums. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology* (pp. 172–176).

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). Arlington. doi: https://doi.org/10.1176/appi.books.9780890425596

Bachmann, S. (2018). Epidemiology of suicide and the psychiatric perspective. *International Journal of Environmental Research and Public Health*, *15*(7), 1–23. doi: 10.3390/ijerph15071425

Balani, S., & De Choudhury, M. (2015). Detecting and Characterizing Mental Health Related Self-Disclosure in Social Media. In *Proceedings of the 33rd annual*

acm conference extended abstracts on human factors in computing systems (pp. 1373–1378). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2702613.2732733

Bayram, U., & Benhiba, L. (2021). Determining a Person's Suicide Risk by Voting on the Short-Term History of Tweets for the CLPsych 2021 Shared Task. In *Proceedings of the seventh workshop on computational linguistics and clinical psychology* (pp. 81–86). Association for Computational Linguistics. doi: 10.18653/v1/2021.clpsych-1.8

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer. *arXiv:2004.05150*.

Benton, A., Coppersmith, G., & Dredze, M. (2017). Ethical Research Protocols for Social Media Health Research. In *Proceedings of the first {acl} workshop on ethics in natural language processing* (pp. 94–102). Valencia, Spain: Association for Computational Linguistics. doi: 10.18653/v1/W17-1612

Benton, A., Mitchell, M., & Hovy, D. (2017). Multi-Task Learning for Mental Health using Social Media Text. *CoRR*, *abs/1712.0*.

Bertolote, J. M., & Fleischmann, A. (2002). Suicide and psychiatric diagnosis: a worldwide perspective. *World psychiatry : official journal of the World Psychiatric Association (WPA)*, *1*(3), 181–185.

Bertolote, J. M., Fleischmann, A., De Leo, D., & Wasserman, D. (2004). Psychiatric diagnoses and suicide: Revisiting the evidence. *Crisis*, *25*(4), 147–155. doi: 10.1027/0227-5910.25.4.147

Bishop, C. (1995). *Neural networks for pattern recognition*. Oxford, U.K.: Oxford University Press.

Bishop, C. (2006). *Pattern Recognition and Machine Learning.*

Bitew, S. K., Bekoulis, G., Deleu, J., Sterckx, L., Zaporojets, K., Demeester, T., & Develder, C. (2019). Predicting Suicide Risk from Online Postings in Reddit The UGent-IDLab submission to the CLPysch 2019 Shared Task A. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology* (pp. 158–161). Minneapolis, Minnesota: Association for Computational Linguistics. doi: 10.18653/v1/W19-3019

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Brådvik, L. (2018). Suicide risk and mental disorders. *International Journal of Environmental Research and Public Health*, *15*(9). doi: 10.3390/ijerph15092028

Brownlee, J. (2018). *Better Deep Learning: Train Faster, Reduce Overfitting, and Make Better Predictions.* (1st ed.). Machine Learning Mastery.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167. doi: 10.1023/A:1009715923555

Burnap, P., Colombo, G., & Scourfield, J. (2015). Machine Classification and Analysis of Suicide-Related Communication on Twitter. In *Ht '15: Proceedings of the 26th acm conference on hypertext & social media* (pp. 75–84). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2700171.2791023

Caruana, R. (1997). Multitask Learning. *Machine Learning*, *28*(1), 41–75. doi: 10.1023/A:1007379606734

Chen, L., Aldayel, A., Bogoychev, N., & Gong, T. (2019). Similar Minds Post Alike:

Assessment of Suicide Risk Using a Hybrid Model. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology* (pp. 152–157). Association for Computational Linguistics. doi: 10.18653/v1/W19-3018

Chollet, F. (2017). *Deep Learning With Python* (1st ed., Vol. 1). Manning Publications. doi: 10.1017/CBO9781107415324.004

Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., & Goharian, N. (2018). SMHD: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1485–1497). Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Colombo, G. B., Burnap, P., Hodorog, A., & Scourfield, J. (2016). Analysing the connectivity and communication of suicidal users on twitter. *Computer Communications*, *73*, 291–300. doi: 10.1016/j.comcom.2015.07.018

Coppersmith, G., Dredze, M., & Harman, C. (2014a). Measuring Post Traumatic Stress Disorder in Twitter. In *International conference on weblogs and social media (icwsm)* (pp. 579–582).

Coppersmith, G., Dredze, M., & Harman, C. (2014b). Quantifying Mental Health Signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 51–60). Baltimore, Maryland, USA: Association for Computational Linguistics. doi: 10.3115/v1/W14-3207

Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. (2015). From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses. In *Proceedings of the 2nd workshop on computational*

linguistics and clinical psychology: From linguistic signal to clinical reality (pp. 1–10). Denver, Colorado: Association for Computational Linguistics. doi: 10.3115/v1/W15-1201

Coppersmith, G., Dredze, M., Harman, C., Kristy, H., & 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). Denver, Colorado: Association for Computational Linguistics. doi: 10.3115/v1/W15-1204

Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural Language Processing of Social Media as Screening for Suicide Risk. *Biomedical Informatics Insights*, *10*. doi: 10.1177/1178222618792860

Coppersmith, G., Ngo, K., Leary, R., & Wood, A. (2016). Exploratory Analysis of Social Media Prior to a Suicide Attempt. In *Proceedings of the 3rd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 106–117). San Diego, CA, USA: Association for Computational Linguistics. doi: 10.18653/v1/W16-0311

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*, 273–297.

De Choudhury, M., Counts, S., & Horvitz, E. (2013a). Major Life Changes and Behavioral Markers in Social Media : Case of Childbirth. In *Computer supported cooperative work (cscw)* (pp. 1431–1442). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2441776.2441937

De Choudhury, M., Counts, S., & Horvitz, E. (2013b). Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the sigchi conference*

on human factors in computing systems - chi '13 (p. 3267–3276). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2470654.2466447

De Choudhury, M., Counts, S., & Horvitz, E. (2013c). Social media as a measurement tool of depression in populations. In *Proceedings of the 5th annual acm web science conference* (pp. 47–56). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2464464.2464480

De Choudhury, M., Counts, S., Horvitz, E. J., & Hoff, A. (2014). Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th acm conference on computer supported cooperative work & social computing (cscw '14)* (pp. 626–638). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2531602.2531675

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016). Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 2098–2110). doi: 10.1145/2858036.2858207

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 4171–4186). Minneapolis, MN, USA: Association for Computational Linguistics. doi: 10.18653/v1/n19-1423

Dome, P., Rihmer, Z., & Gonda, X. (2019). Suicide risk in bipolar disorder: A brief review. *Medicina (Lithuania)*, *55*(8). doi: 10.3390/medicina55080403

Dubé, D.-E. (2020). *University of Ottawa confirms student death on-campus.*

Retrieved from https://ottawa.citynews.ca/local-news/university-of-ottawa-confirms-student-death-on-campus-2166694

Fast, E., Chen, B., & Bernstein, M. (2016). Empath: Understanding Topic Signals in Large-Scale Text. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 4647–4657). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2858036.2858535

Gamoran, A., Kaplan, Y., Simchon, A., & Gilead, M. (2021). Using Psychologically-Informed Priors for Suicide Prediction in the CLPsych 2021 Shared Task. In Online (Ed.), *Proceedings of the seventh workshop on computational linguistics and clinical psychology* (pp. 103–109). Association for Computational Linguistics. doi: 10.18653/v1/2021.clpsych-1.12

Gaur, M., Kursuncu, U., Sheth, A., Alambo, A., Thirunarayan, K., Welton, R. S., ... Pathak, J. (2019). Knowledge-aware assessment of severity of suicide risk for early intervention. In *The world wide web conference* (pp. 514–525). doi: 10.1145/3308558.3313698

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. Retrieved from http://www.deeplearningbook.org

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the eleventh international conference on language resources and evaluation 2018*. Miyazaki, Japan: European Language Resources Association (ELRA).

Handley, T., Rich, J., Davies, K., Lewin, T., & Kelly, B. (2018). The challenges of predicting suicidal thoughts and behaviours in a sample of rural Australians with depression. *International Journal of Environmental Research and Public*

Health, *15*(5). doi: 10.3390/ijerph15050928

Harrigian, K., Aguirre, C., & Dredze, M. (2020). Do Models of Mental Health Based on Social Media Data Generalize? In *Findings of the association for computational linguistics: Emnlp 2020* (pp. 3774–3788). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.337

Hawton, K., Comabella, C. C. I., Haw, C., & Saunders, K. (2013). Risk factors for suicide in individuals with depression: A systematic review. *Journal of Affective Disorders*, *147*(1-3), 17–28. doi: 10.1016/j.jad.2013.01.004

Heinzerling, B., & Strube, M. (2018). BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In *Proceedings of the eleventh international conference on language resources and evaluation 2018.* Miyazaki, Japan: European Language Resources Association (ELRA).

Holmstrand, C., Bogren, M., Mattisson, C., & Brådvik, L. (2015). Long-term suicide risk in no, one or more mental disorders: The Lundby Study 1947-1997. *Acta Psychiatrica Scandinavica*, *132*(6), 459–469. doi: 10.1111/acps.12506

Huang, X., Li, X., Zhang, L., Liu, T., Chiu, D., & Zhu, T. (2015). Topic Model for Identifying Suicidal Ideation in Chinese Microblog. In *29th pacific asia conference on language, information and computation* (pp. 553–562). Shanghai.

Husseini Orabi, A., Buddhitha, P., Husseini Orabi, M., & Inkpen, D. (2018). Deep Learning for Depression Detection of Twitter Users. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: From keyboard to clinic* (pp. 88–97). New Orleans, LA: Association for Computational Linguistics. doi: 10.18653/v1/w18-0609

Iserman, M., Nalabandian, T., & Ireland, M. (2019). Dictionaries and Decision Trees

for the 2019 CLPsych Shared Task. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology* (pp. 188–194). Minneapolis, Minnesota: Association for Computational Linguistics. doi: 10.18653/v1/w19-3025

Jamison-Powell, S., Linehan, C., Daley, L., Garbett, A., & Lawson, S. (2012). I can't get no sleep: discussing #insomnia on Twitter. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1501–1510). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2207676.2208612

Jaroszewski, A. C., Morris, R. R., & Nock, M. K. (2019). Randomized controlled trial of an online machine learning-driven risk assessment and intervention platform for increasing the use of crisis services. *Journal of Consulting and Clinical Psychology*, *87*(4), 370–379. doi: 10.1037/ccp0000389

Ji, S., Li, X., Huang, Z., & Cambria, E. (2021). Suicidal Ideation and Mental Disorder Detection with Attentive Relation Networks. *Neural Computing and Applications*, *0123456789*. doi: 10.1007/s00521-021-06208-y

Joinson, A. N., & Paine, C. B. (2007). Self-disclosure, Privacy and the Internet. *The Oxford handbook of Internet psychology*, *2374252*, 237–252. doi: 10.1093/oxfordhb/9780199561803.013.0016

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics* (pp. 427–431). Valencia, Spain: Association for Computational Linguistics.

Kessler, J. S. (2017). ScatterText: A browser-based tool for visualizing how corpora differ. In *Proceedings of acl 2017, system demonstrations* (pp. 85–90). Vancouver,

Canada: Association for Computational Linguistics.

Kim, S. M., Wang, Y., & Wan, S. (2016). Data61-CSIRO systems at the CLPsych 2016 Shared Task. In *Proceedings of the third workshop on computational linguistics and clinical psycholog* (pp. 128–132). San Diego, CA, USA: Association for Computational Linguistics. doi: 10.18653/v1/W16-0313

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1746–1751). Doha, Qatar: Association for Computational Linguistics. doi: 10.3115/v1/d14-1181

Kingma, D. P., & Ba, J. L. (2015). Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd international conference on learning representations 2015* (pp. 1–15). San Diego, CA, USA.

Kitaev, N., Kaiser, L., & Levskaya, A. (2020). Reformer: The Efficient Transformer. In *International conference on learning representations.*

Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. (2017). Self-normalizing neural networks. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 972–981). Red Hook, NY, USA: Curran Associates Inc.

Kshirsagar, R., Morris, R., & Bowman, S. (2017). Detecting and Explaining Crisis. In *Proceedings of the fourth workshop on computational linguistics and clinical psychology — from linguistic signal to clinical reality* (pp. 66–73). Vancouver, BC: Association for Computational Linguistics. doi: 10.18653/v1/W17-3108

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016*

conference of the north american chapter of the association for computational linguistics: Human language technologies (pp. 260–270). San Diego, California: Association for Computational Linguistics. doi: 10.18653/v1/N16-1030

LeBouthillier, D. M., McMillan, K. A., Thibodeau, M. A., & Asmundson, G. J. G. (2015). Types and Number of Traumas Associated With Suicidal Ideation and Suicide Attempts in PTSD: Findings From a U.S. Nationally Representative Sample. *Journal of Traumatic Stress*, *28*(3), 183–190. doi: 10.1002/jts.22010

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural computation*, *1*(4), 541–551. doi: 10.1162/neco.1989.1.4.541

Lehrman, M. T., Alm, C. O., & Proaño, R. A. (2012). Detecting Distressed and Non-distressed Affect States in Short Forum Texts. In *Proceedings of the second workshop on language in social media* (pp. 9–18). Montréal, Canada: Association for Computational Linguistics.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2018). Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, *18*, 1–52.

Linthicum, K. P., Schafer, K. M., & Ribeiro, J. D. (2019). Machine learning in suicide science: Applications and ethics. *Behavioral Sciences and the Law*, *37*(3), 214–222. doi: 10.1002/bsl.2392

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loveys, K., Crutchley, P., Wyatt, E., & Coppersmith, G. (2017). Small but Mighty: Affective Micropatterns for Quantifying Mental Health from Social Media Language. In *Proceedings of the fourth workshop on computational linguistics and clinical psychology, from linguistic signal to clinical reality* (pp. 85–95). Vancouver, BC: Association for Computational Linguistics. doi: 10.18653/v1/W17-3110

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the 30th international conference on machine learning* (Vol. 28). Atlanta, Georgia, USA.

Macavaney, S., Desmet, B., Cohan, A., Soldaini, L., Yates, A., Zirikly, A., & Goharian, N. (2018). RSDD-Time : Temporal Annotation of Self-Reported Mental Health Diagnoses. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: From keyboard to clinic* (pp. 168–173). New Orleans, LA: Association for Computational Linguistics. doi: 10.18653/v1/W18-0618

MacAvaney, S., Mittu, A., Coppersmith, G., Leintz, J., & Resnik, P. (2021). Community-level Research on Suicidality Prediction in a Secure Environment: Overview of the CLPsych 2021 Shared Task. In *Proceedings of the seventh workshop on computational linguistics and clinical psychology* (pp. 70–80). Online: Association for Computational Linguistics. doi: 10.18653/v1/2021.clpsych-1.7

Malmasi, S., Zampieri, M., & Dras, M. (2016). Predicting Post Severity in Mental Health Forums. In *Proceedings of the third workshop on computational linguistics and clinical psycholog* (pp. 133–137). San Diego, CA, USA: Association for Computational Linguistics. doi: 10.18653/v1/W16-0314

Masters, D., & Luschi, C. (2018). Revisiting Small Batch Training for Deep Neural Networks. *CoRR*, *abs/1804.0*.

Matero, M., Idnani, A., Son, Y., Giorgi, S., Vu, H., Zamani, M., ... Schwartz, H. A. (2019). Suicide Risk Assessment with Multi-level Dual-Context Language and. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology* (pp. 39–44). Minneapolis, Minnesota: Association for Computational Linguistics. doi: 10.18653/v1/W19-3005

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the 26th international conference on neural information processing systems* (p. 3111–3119). Red Hook, NY, USA: Curran Associates Inc.

Milne, D. N., McCabe, K. L., & Calvo, R. A. (2019). Improving moderator responsiveness in online peer support through automated triage. *Journal of Medical Internet Research*, *21*(4). doi: 10.2196/11410

Mitchell, M., Hollingshead, K., & Coppersmith, G. (2015). Quantifying the Language of Schizophrenia in Social Media. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 11–20). Denver, Colorado: Association for Computational Linguistics. doi: 10.3115/v1/W15-1202

Mohammad, S. M., & Bravo-Marquez, F. (2017). WASSA-2017 Shared Task on Emotion Intensity. In *Proceedings of the workshop on computational approaches to subjectivity, sentiment and social media analysis (wassa).* Copenhagen, Denmark.

Mohammadi, E., Amini, H., & Kosseim, L. (2019). CLaC at CLPsych 2019: Fusion of Neural Features and Predicted Class Probabilities for Suicide Risk Assessment Based on Online Posts. In *Proceedings of the sixth workshop on computational*

linguistics and clinical psychology (pp. 34–38). Association for Computational Linguistics. doi: 10.18653/v1/W19-3004

Morales, M., Dey, P., & Kohli, K. (2021). Team 9: A Comparison of Simple vs. Complex Models for Suicide Risk Assessment. In *Proceedings of the seventh workshop on computational linguistics and clinical psychology: Improving access* (pp. 99–102). Online: Association for Computational Linguistics. doi: 10.18653/v1/2021.clpsych-1.11

Morales, M., Dey, P., Theisen, T., Belitz, D., & Chernova, N. (2019). An Investigation of Deep Learning Systems for Suicide Risk Assessment. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology* (pp. 177–181). Minneapolis, Minnesota: Association for Computational Linguistics. doi: 10.18653/v1/W19-3023

Nair, V., & Hinton, G. E. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th international conference on international conference on machine learning* (p. 807–814). Madison, WI, USA: Omnipress.

Nobles, A. L., Glenn, J. J., Kowsari, K., Teachman, B. A., & Barnes, L. E. (2018). Identification of Imminent Suicide Risk Among Young Adults using Text Messages. In *Proceedings of the sigchi conference on human factors in computing systems. chi conference* (pp. 1–11). doi: 10.1145/3173574.3173987

Nock, M. K., Hwang, I., Sampson, N., Kessler, R. C., Angermeyer, M., Beautrais, A., ... Williams, D. R. (2009). Cross-national analysis of the associations among mental disorders and suicidal behavior: Findings from the WHO World Mental Health Surveys. *PLoS Medicine*, *6*(8). doi: 10.1371/journal.pmed.1000123

Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., & Dehak, N. (2019, 12). Hierarchical Transformers for Long Document Classification. In *2019 ieee automatic speech recognition and understanding workshop (asru)* (pp. 838–844). IEEE. doi: 10.1109/ASRU46091.2019.9003958

Park, M., Mcdonald, D. W., & Cha, M. (2013). Perception Differences between the Depressed and Non-depressed Users in Twitter. In *Proceedings of the 7th international aaai conference on weblogs and social media (icwsm)* (pp. 476–485). Cambridge, MA, United States.

Pavalanathan, U., & De Choudhury, M. (2015). Identity Management and Mental Health Discourse in Social Media. In *Proceedings of the international world-wide web conference. international www conference* (pp. 315–321). doi: 10.1145/2740908.2743049

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The Development and Psychometric Properties of LIWC2007 The University of Texas at Austin. *Austin, TX, LIWC. Net*. doi: 10.1068/d010163

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics. doi: 10.3115/v1/D14-1162

Preot, D., Eichstaedt, J., Park, G., Sap, M., Smith, L., Tobolsky, V., ... Ungar, L. (2015). The Role of Personality , Age and Gender in Tweeting about

Mental Illnesses. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 21–30). Denver, Colorado: Association for Computational Linguistics. doi: 10.3115/v1/W15-1203

Preotiuc-Pietro, D., Sap, M., Schwartz, H. A., & Ungar, L. (2015). Mental Illness Detection at the World Well-Being Project for the CLPsych 2015 Shared Task. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 40–45).

Raaijmakers, S. (2021). *Deep Learning for Natural Language Processing*. Manning Publications.

Resnik, P., Armstrong, W., Claudino, L., & Nguyen, T. (2015). The University of Maryland CLPsych 2015 Shared Task System. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 54–60). Denver, Colorado: Association for Computational Linguistics. doi: 10.3115/v1/W15-1207

Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.-A., & Boyd-graber, J. (2015). Beyond LDA : Exploring Supervised Topic Modeling for Depression-Related Language in Twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 99–107). Denver, Colorado: Association for Computational Linguistics. doi: 10.3115/v1/W15-1212

Resnik, P., Foreman, A., Kuchuk, M., Musacchio Schafer, K., & Pinkham, B. (2021, 2). Naturally occurring language as a source of evidence in suicide prevention. *Suicide and Life Threatening Behavior*, *51*(1), 88–96. doi: 10.1111/sltb.12674

Resnik, P., Garron, A., & Resnik, R. (2013). Using Topic Modeling to Improve Prediction of Neuroticism and Depression in College Students. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1348–1353). Seattle, Washington, USA: Association for Computational Linguistics.

Ruder, S. (2017, 6). An Overview of Multi-Task Learning in Deep Neural Networks. *CoRR*, *abs/1706.0*.

Ruiz, V., Shi, L., Quan, W., Ryan, N., Biernesser, C., Brent, D., & Tsui, R. (2019). CLPsych2019 Shared Task: Predicting Suicide Risk Level from Reddit Posts on Multiple Forums. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology* (pp. 162–166). Minneapolis, Minnesota: Association for Computational Linguistics. doi: 10.18653/v1/W19-3020

Samaritans. (2015). *Samaritans Radar.* Retrieved from `https://www.samaritans.org/about-samaritans/research-policy/internet-suicide/samaritans-radar/`

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020, 10). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, 2–6.

Schimmele, C., Fonberg, J., & Schellenberg, G. (2021). *Canadians' assessments of social media in their lives* (Tech. Rep.). Statistics Canada. Retrieved from `https://www150.statcan.gc.ca/n1/pub/36-28-0001/2021003/article/00004-eng.htm`

Schwartz, H. A., Eichstaedt, J., Kern, M. L., Park, G., Sap, M., Stillwell, D., ... Ungar, L. (2014). Towards Assessing Changes in Degree of Depression through Facebook.

In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 118–125). Baltimore, Maryland, USA: Association for Computational Linguistics. doi: 10.3115/v1/W14-3214

Shing, H.-c., Nair, S., Zirikly, A., Friedenberg, M., Daumé III, H., & Resnik, P. (2018). Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: From keyboard to clinic* (pp. 25–36). New Orleans, LA: Association for Computational Linguistics. doi: 10.18653/v1/W18-0603

Simpson, S. G., & Jamison, K. R. (1999). The risk of suicide in patients with bipolar disorders. *The Journal of clinical psychiatry*, *60 Suppl 2*, 53–6.

Skenderi, S. (2016). *Experts caution about use of unmonitored mental health app forums.* Retrieved from `https://toronto.citynews.ca/2016/04/29/experts-caution-about-use-of-unmonitored-mental-health-app-forums/`

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*(2), 1929–1958.

Statistics Canada. (2019). *Mental health care needs* (Tech. Rep.). Retrieved from `https://www150.statcan.gc.ca/n1/pub/82-625-x/2019001/article/00011-eng.pdf`

Statistics Canada. (2020). *Canadian Community Health Survey, 2019* (Tech. Rep.). Ottawa: Statistics Canada. Retrieved from `https://www150.statcan.gc.ca/n1/en/daily-quotidien/200806/dq200806a-eng.pdf?st=YfuWAj78`

Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019, 12). Detection of Suicide Ideation in Social Media Forums Using Deep Learning. *Algorithms*, *13*(1), 7.

doi: 10.3390/a13010007

Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., & Ohsaki, H. (2015, 4). Recognizing Depression from Twitter Activity. In *Proceedings of the 33rd annual acm conference on human factors in computing systems* (pp. 3187–3196). New York, NY, USA: ACM. doi: 10.1145/2702123.2702280

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017, 11). Attention is All you Need. In *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

Wang, N., Fan, L., Shivtare, Y., Badal, V., Subbalakshmi, K., Chandramouli, R., & Lee, E. (2021). Learning Models for Suicide Prediction from Social Media Posts. In *Proceedings of the seventh workshop on computational linguistics and clinical psychology: Improving access* (pp. 87–92). Online: Association for Computational Linguistics. doi: 10.18653/v1/2021.clpsych-1.9

Wilcox, H. C., Storr, C. L., & Breslau, N. (2009, 3). Posttraumatic Stress Disorder and Suicide Attempts in a Community Sample of Urban American Young Adults. *Archives of General Psychiatry*, *66*(3), 305–311. doi: 10.1001/archgenpsychiatry.2008.557

World Health Organization. (2004). *Prevention of mental disorders* (Tech. Rep.). Geneva. Retrieved from `https://www.who.int/mental_health/evidence/en/prevention_of_mental_disorders_sr.pdf`

World Health Organization. (2018). *Mental disorders.* Retrieved from `https://www.who.int/en/news-room/fact-sheets/detail/mental-disorders`

World Health Organization. (2021). *Suicide worldwide in 2019: global health estimates* (Tech. Rep.). Geneva: World Health Organization. Retrieved from `https://`

www.who.int/teams/mental-health-and-substance-use/suicide-data

Wu, S., Zhang, H. R., & Ré, C. (2020). Understanding and Improving Information Transfer in Multi-Task Learning. In *International conference on learning representations.*

Xu, B., Wang, N., Chen, T., & Li, M. (2015). Empirical Evaluation of Rectified Activations in Convolutional Network. *ArXiv*, *abs/1505.0*.

Yogaretnam, S. (2020). *Five student deaths in 10 months: UOttawa faces mental health crisis.* Retrieved from https://ottawacitizen.com/news/five-student-deaths-in-10-months-uottawa-faces-mental-health-crisis

Zaheer, J., Olfson, M., Mallia, E., Lam, J. S., de Oliveira, C., Rudoler, D., ... Kurdyak, P. (2020, 8). Predictors of suicide at time of diagnosis in schizophrenia spectrum disorder: A 20-year total population study in Ontario, Canada. *Schizophrenia Research*, *222*, 382–388. doi: 10.1016/j.schres.2020.04.025

Zirikly, A., Resnik, P., Uzuner, O., & Hollingshead, K. (2019). CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology* (pp. 24–33). Minneapolis, Minnesota: Association for Computational Linguistics. doi: 10.18653/v1/W19-3003