UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Deep Learning-Based Frameworks for Automated Identifying Depression Through Social Media

A thesis submitted in partial fulfilment of the requirements
for the degree of

Doctor of Philosophy
in
Analytics

by

Hamad Zogan

to

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2007, Australia

April 2023

CERTIFICATE OF ORIGINAL AUTHORSHIP

I *Hamad Zogan* declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Computer Science/Faculty of Engineering and Information Technology at the University of Technology Sydney, Australia.
This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

SIGNATURE: _____

DATE: 24th April, 2023

Abstract

Twitter is a widely used social media website that allows individuals to share their own content with the public. The data generated by users on this platform is extremely valuable for healthcare technology as it can reveal important patterns that can greatly benefit the field in multiple ways. An example application is the automatic identification of mental health issues such as depression. Notably, the existing research on online depression detection is limited, with main challenges. First, previous research on identifying depressed users on social media primarily focused on analyzing user behavior and language patterns, including their social interactions. However, these methods have a drawback as they tend to be trained on irrelevant content that may not be essential for detecting depression, which can negatively impact the model's efficiency and effectiveness. Second, limited research has been conducted to identify the changes and variations in depression levels at a more specific level, such as state or neighborhood level, during the COVID-19 pandemic. Third, the ability to explain a model's predictions is crucial for gaining trust, as it offers an understanding of how the model came to a certain conclusion. Unfortunately, many machine learning techniques lack explainability, which is a concern. For example, in the task of automatically predicting depression, most machine learning models produce predictions that are not easily understandable to humans. Fourth, creating new tasks aimed specifically towards modeling narrative elements in social media and to what extent using social media posts makes it possible to extract such features for a narrative explanation of a series of events, which could be significant if we compare people with a mental disorder. Therefore, this thesis aims to develop approaches to identifying depression using online social media, particularly Twitter, and build prediction models that can identify users who are likely to be experiencing mental problems or displaying symptoms that might soon lead to mental disorders. This thesis is organized into five main themes: (1) A depression classification model for understanding how the COVID-19 pandemic has affected people's depression; (2) Depression detection at the User level and its impact during the pandemic; (3) A new, scalable hybrid model that utilizes a combination of deep learning techniques to identify depressed individuals on social media platforms like Twitter through the use of multiple features; (4) An explicable deep learning-based system for depression detection; (5) Modeling narrative elements to identify depression. This thesis's empirical results and findings show the advantages of the proposed approaches in that achieve outstanding performance and provide sufficient evidence to justify the predictions, and demonstrate the narrative elements of a depressed user.

Keywords: Depression Detection, Social Network, Deep Learning, Explainability, COVID-19, Twitter, Australia

ACKNOWLEDGMENTS

I would like to express my sincere gratitude and acknowledge the financial support provided by Jazan University, which has enabled me to pursue my academic studies at the University of Technology Sydney. This support is highly acknowledged and appreciated. Moreover, I am thankful to my supervisor Prof. Guandong Xu, Dr Muhammad Imran Razzak and Dr Shoaib Jamee for all of their advice, guidance and support during my PhD degree course at the University of Technology Sydney. This work couldn't have been done without their professional direction and support.

In conclusion, I would like to express my heartfelt appreciation to my wife, Annalisa Pirrello, for her unwavering love, support, and encouragement throughout my academic journey. Her patience, understanding, and motivation have been vital to my success in completing this thesis. Additionally, I am deeply grateful to my family for their constant support and encouragement, which have been integral to my research studies. Their love and encouragement have sustained me during the long and arduous process of writing this thesis, and I could not have accomplished this feat without them.

List of Publications

List of Journal Articles (Published)

1. Jianlong Zhou, **Hamad Zogan**, Shuiqiao Yang, Shoaib Jameel, Guandong Xu and Fang Chen. "Detecting community depression dynamics due to covid-19 pandemic in Australia." **IEEE Transactions on Computational Social Systems**, (2021).

2. **Hamad Zogan**, Imran Razzak, Xianzhi Wang, Shoaib Jameel and Guandong Xu. "Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media." **World Wide Web Journal**, (2022).

3. **Hamad Zogan**, Imran Razzak, Shoaib Jameel and Guandong Xu. "Convolutional Attention Based Multimodal Network for Depression Detection on Social Media and Its Impact During Pandemic." **Journal of Biomedical and Health Informatics**, (2023).

List of Conference Articles (Published)

1. **Hamad Zogan**, Imran Razzak, Shoaib Jameel and Guandong Xu. "Depressionnet: learning multi-modalities with user post summarization for depression detection on social media". **SIGIR '21**, (2021).

Table of Contents

List of Figures

List of Tables

Table **Page**

CHAPTER 1

INTRODUCTION

This chapter presents the background of depression, depression in social media, and the motivation of this research, followed by the thesis objectives, contributions, and the organization of this thesis to present the research workflow are outlined and discussed.

1.1 Background

1.1.1 Major Depression Disorder

Depression is a mental condition described by a constant sense of melancholy and disinterest. It differs from the mood swings that people often encounter as part of daily living. There are several causes of depression. Events in life might start depressive episodes or make them worse in the case of unfavorable living conditions. Depression is often exacerbated by a sense of unworthiness and pessimism about oneself, and the world (2). Numerous forms of abuse, unemployment, sexual orientation, financial status, the dissolution of personal relationships, and a sense of loneliness are some of the most frequent reasons for depression. Nearly 264 million people, or 3.4% of the world's population, suffer from diseases linked to depression, according to a World Health Organization research (73). Adults between the ages of 15 and 29 have a greater risk of depression. Every day, this rate is rising. Suicide rates have also grown to be a significant worry as depression rates rise (199). Moreover, depression is one of the most common types of mental disorder in our contemporary society, which is characterized by a rapid

reliance on technological advancement. The current methods fall short of meeting the general public's collective demands for mental health treatment, even on the most fundamental level. It is crucial to remember that there is currently no known cure for this sickness that can reverse all of its symptoms. Therefore, in order to prevent the issue from getting significantly worse over time, it is crucial that we identify its root causes and find a solution.

1.1.2 Depression on Social Media

Usually, it can be difficult to identify depression on a broad scale since the majority of conventional methods of diagnosis rely on questionnaires, interviews, self-reports, or evidence from family and friends. These techniques are not scalable, which would not enable them to reach a more significant population. In order to scale their method to some level and reach more impacted groups in less time, the conventional contact between people and health organizations has changed. They now meet online by sharing information in online communities, seeking guidance, and providing advice. Recent studies show that many users on social media prefer to post or provide advice on health-related information in addition to discussing their moods and behaviors. (60; 116; 153; 137). These resources offer a possible route for learning about mental health for tasks including diagnosis, treatment, and claims. Over the past ten years, several attempts have been made to evaluate non-clinical data in order to evaluate associated symptoms, such as depressive disorders, self-harm, and the severity of mental illness (19). Due to the enormous amount of data, social media platforms and other online discussion forums have attracted the research community's attention for various study goals (such as population-level mental health monitoring (31), cyberbullying detection (19), etc.). Due to rising internet usage and people's impulsive, anonymous sharing of their suffering on these platforms, there has been a tremendous inflow of data (124).

Social networking is a great tool for automatically identifying those who are depressed. While it would take a long time to manually go through every social media post and profile to find persons experiencing depression, scalable computer approaches might offer quick and widespread identification of depressed people, which might aid those who need help at the proper time and avert many catastrophic tragedies in the future. Data miners may find a gold mine in the daily activities of social media users since this information

[1] https://https://www.statista.com/chart/19262/impact-of-social-media-on-mental-health/

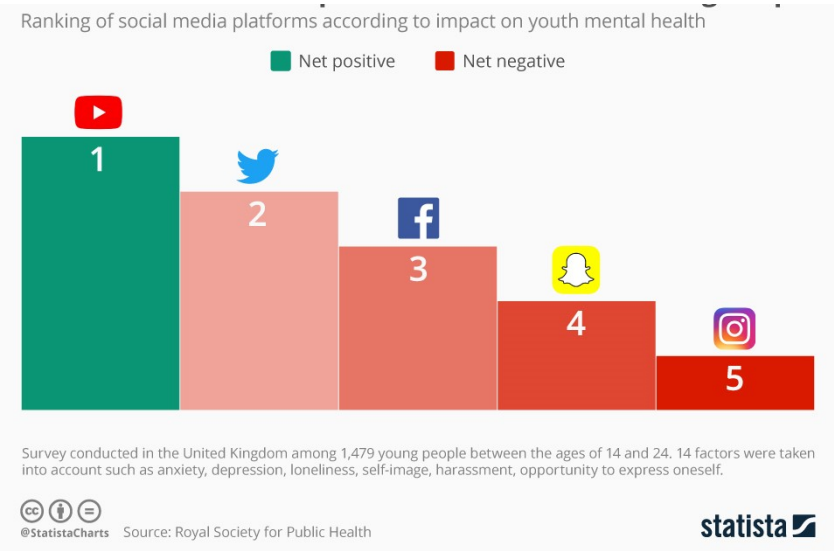


Figure 1.1: Mental health: The impact of Social media on Young People[1].

helps them get deep insights into user-generated content. In addition to giving them a new platform to investigate user behavior, it also makes it feasible to analyze fascinating data in ways that would not otherwise be possible. By analyzing users' online posting behaviors on various social networks like Facebook, Weibo (102; 42), Twitter, and others (see Fig.1.1), psychologists and scientists may learn about users' behavioral patterns to better target the right individuals with the appropriate care at the right time (22).

The vast volume and rapidity of media content posted on the web represent a huge challenge for human analysts. The increasing number of users expressing their opinions in real-time is generating a massive amount of data that provides a treasure trove of information, which can be used in various fields, such as disaster management (45; 1), news (152; 149; 18), sports (216), and traffic.(130), etc. Microblogging sites like Twitter, Facebook, etc., have become essential platforms for people to communicate. It allows users to post short messages to their audience over the Internet. These messages are a mixture of microblogging and micro-messages that consist of texts, pictures or videos. These platforms are turning into the main source for people to uncover their daily feelings and psychological problems. Twitter, in particular, stands out as a public platform due to its unique way of letting people view and express every bit of essential events. Twitter is a space for people to express their opinions on many topics and people to share their life experiences. It grows excessively and generates a tremendous amount of data. Twitter today, is one of the fastest and most popular platforms of communication with thousands of people tweeting daily about their daily life events. Healthcare took advantage of

the vast amount of data on social media to find hidden insights. However, researchers who apply data analysis to social network data should be aware of various challenges, including differentiating between sentiment and general data, quality (how meaningful certain messages and comments are to certain individuals), temporal relevance (what is relevant today may not be relevant tomorrow), and how viral activity begins and spreads on social media platforms. However, there could be significant challenges in trying to classify every user post available in the profile due to the number of tweets for each user, e.g., the large and diverse number of posts per user. One of the problems is dealing with the "curse-of-dimensionality" problem which might have an overall deteriorating impact on the performance in terms of time and space complexities of the model.

Twitter has recently been a more effective method for identifying mental illnesses like depression in several parts of the world. On Twitter, vocabulary usage has been proven to signify depression. For instance, it was discovered that depressed users used verbs more frequently and tended to use the first-person singular pronoun the most (87). As a result, much research has been done to apply machine learning approaches to extract data such as users' social activity habits, user profiles, and texts from their social media posts in order to detect depression(36; 178; 208; 71; 156).

Additionally, because of the lockdown in impacted regions and reports of greater rates of mood disorders, including acute stress disorder, and PTSD, the COVID-19 pandemic breakout is anticipated to have severe repercussions on the mental health of millions of people. An overall decline in subclinical mental health, as well as stress disorder and generalized anxiety disorder (165). The COVID-19 pandemic's severity of mental health deterioration and the thoroughness of detecting depressive illnesses have created an unprecedented necessity to infer people's mental states from extensive sources. It is now more important than ever to infer people's mental states from a variety of sources due to the severity of mental health decline brought on by the COVID-19 pandemic and the completeness of diagnosing depressive diseases. Posts or comments on social media can offer crucial information on how the pandemic impacts mental health in the community, according to recent research (99).

Moreover, previous studies have looked into the possibility of extracting information on mental health disorders from social media data, either by aggregating user activity patterns or by utilizing regular expressions to find self-reported diagnoses. However, The accessibility of massive, high-quality annotated datasets that address the severity of the mental sickness is a significant progression component. Unfortunately, there are not many datasets available for the severity of depression that also do not include

trustworthy baseline information based on clinical validation (174).

1.2 Motivation

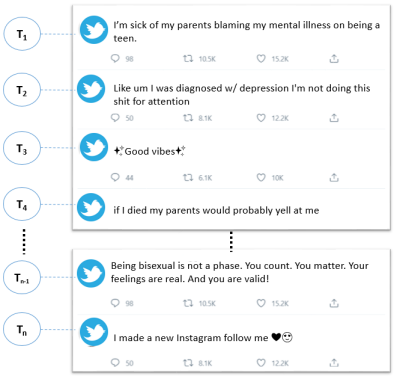


Figure 1.2: A sample of depressed user tweets.

Most of the recent depression detection models are limited to detecting in a large number of posts; therefore, one of the major shortcomings underlying existing methods conducting depression detection on social media is using both depression-focused and depression-irrelevant posts. It is common for a user to share different kinds of posts online on various topics conveying different opinions and interests. Some of these posts may contain no discriminative information on the depressed state of the user. Co-occurring frequency of terms between user posts is low because twitter's post is very short with diverse topics, limiting depression detection significantly. We show this through the example in Figure 1.2 that shows some tweets represent some posts history of a depressed user. Notice that some of these tweets are not relevant to depression mood or symptoms, and considering these irrelevant tweets can heavily impact the identification performance. There could be significant challenges in trying to classify every user post available in the profile due to the number of tweets for each user, e.g., the large and diverse number of posts per user. One of the problems is dealing with the "curse-of-dimensionality" problem, which might have an overall deteriorating impact on the performance in terms of time and space complexities of the model.

Since mental illness effects on social media are perceptible (129), different methods to detect them have been designed In recent years, and it has been studied to detect depression via social media, showing some encouraging early results. In (85) the authors provide a methodology to identify users who are anxious or depressed. They have suggested an ensemble classification model that integrates findings from three widely used

models as well as individual performance analysis of each model in the ensemble. The authors developed a rapid approach to get their dataset by selecting the first 100 randomly selected users who have followed the MS India Student Forum for a month in order to gain relevant data. Many studies have been conducted to employ machine learning approaches to extract data such as users' social activity behaviors, user profiles, and texts from their social media posts in order to detect depression(36; 178; 208; 71; 156). For instance, De Choudhury et al. (36) suggested utilizing support vector machines (SVM) to predict depression for social media users based on Twitter using prediction based on manually labeled training data. Furthermore, building a method that can accurately evaluate tweets containing self-reported depression-related attributes will help people and medical professionals better comprehend the severity of users' depression. Unsupervised strategies to automatically identify depression online have been proposed by researchers (213; 178).

1.3 Research Objectives

This study aims to develop approaches to identifying depression or mental disorders in general, using online social media, particularly Twitter, and build prediction models that can identify users who are likely to be experiencing mental problems or displaying symptoms that might soon lead to mental disorders. Hence, the following research objectives are presented as the central targets to achieve the Research Questions of identifying mental illness through social media:

1. Integrate user posts with his behavior to create a wide range of behavioral, lexical, and semantic representations of users.

2. To automatically choose the most significant user-generated information as a natural boost to our computational architecture for claim analysis and depression detection management.

3. In this study, explainable deep learning architecture is designed and implemented to identify depression in social media.

4. To create new tasks that aim to model user narrative in Social Media that help to understand crucial narrative features and how they evolve and extract such features for A narrative explanation of a series of events.

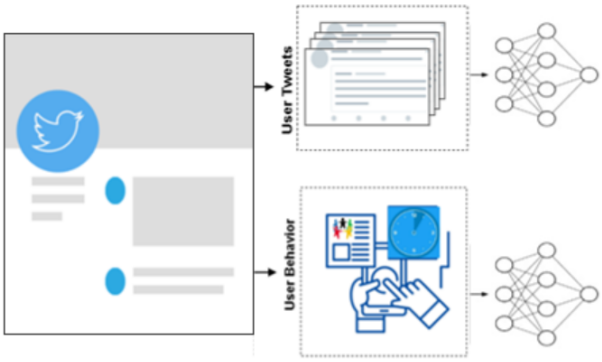


Figure 1.3: Two pieces of information from a user we will analyze user tweets and user behaviors.

These research objectives are the primary motivation for this thesis and related online depression detection research. Moreover, this thesis aims to develop novel techniques for addressing the challenges of online depression detection. The research goal of this thesis is to analyze the characteristics of depression users in social networks comprehensively. Furthermore, find a pattern to identify a depressed user in an early stage. Specifically, We mainly analyze two pieces of information from a user, his behaviors and his timeline tweets (Figure 1.3) to develop approaches to identifying depression. Since the previous study for depression at the user level focused on user behaviors only; therefore, we will study combining user behavior and user post history or user activity. Then We will dig deeper into a user post to propose an explainable depression detection model that can extract the most crucial tweets that could hint at causing a user depression. Since a user's tweets could be crucial and essential, our approach will be to analyze a user's tweets alone, and we will propose new research that can create new specific feature representations and achieve better outcomes for identifying the level of depression in Twitter users. Additionally, we will extract narrative explanations from the content of the user's posts.

1.4 Research Problems and Contributions

In order to achieve the research objectives mentioned in the last section, in the works of this thesis, we aim to investigate the following research problems and propose the following categories of models and approaches.

1.4.1 Learning Multi-Modalities with User Post Summarization

People who suffer from depression sometimes have a tendency to conceal their symptoms in order to keep their condition a secret or find it challenging to contact a professional diagnostician (146; 162). It would take much effort to manually search for such persons online by reading through their tweets. The tendency for existing systems to use each and every social media post made by a user is one of their flaws. We believe that this is unnecessary since it often renders the automatic depression detection system ineffective and even performs worse, such as when dealing with the "curse of dimensionality," and that these irrelevant posts may have a more dominant effect than depression-relevant content. Therefore, this study attempts to detect depression through all user posts based on the questions as follows:

- What is an effective strategy through which we can select features so that we can identify patterns from implicit or passive users?
- Can we provide a computational framework that uses the information acquired after doing content summarization to train and aids in choosing the classifier's most beneficial features?
- Does an abstractive-extractive automated text summarising model that reduces a vast collection of tweets into a succinct, conclusive summary meet full coverage of tweets related to depression?
- Is it sufficient to preserve content that could be related to depression using an abstractive-extractive automated text summarization model?
- How can information about user behaviors on social media be used to make our model prediction more accurate?

For this problem, and in order to aid in choosing the most beneficial features for the classifier, we developed a novel computational framework that was trained on the content acquired following extractive-abstract user-generated content summarization. We have proposed a model that integrates user tweets and various completely connected layers to depict user behavior. Our model utilized a novel summarization framework by using relevant user post history to automatically choose the most important user-generated information. Our computational framework benefits naturally from automatic summarization since it may concentrate only on the essential tweets during model

training. Another advantage of automatic summarization is that our feature selection is not driven by arbitrary design decisions.

1.4.2 Explainability for Depression Detection

Recently, deep learning was used to detect depression on social media, and the results were notably better than those of conventional machine learning techniques. Despite recent research demonstrating the efficacy of deep learning techniques for depression detection, most currently used machine learning techniques do not offer any depression prediction explainability; hence, this makes their predictions difficult for people to understand, which lowers confidence in deep learning algorithms. An explainable model aids understanding and provides guidance on how a deep learning model may improve. Thus, this study attempts to suggest explainable depression detection based on the questions as follows:

- Can we build a deep learning-based depression detection system that explains how decisions are made and why an individual user is depressed?
- How should the explainability of the model for depression detection be presented?
- How can the prediction of depression detection be explained using a pipeline supported by explainability and based on hierarchical attention networks?

To address these concerns, we plan to develop an explainable deep learning-based solution for depression detection by utilizing various features from the different behaviors of depressed users on social media. This model aims to improve the accuracy of identifying depression in Twitter users and extracting explanations from the information they post. Additionally, the model will create new specific feature representations from the training data.

1.4.3 Modeling Narrative Elements to Identify Depression

Furthermore, On social media, people usually engage in events and pursuits that they find meaningful and interesting. Sometimes, the posts about a particular event could be a chronology feed. A chronological feed essentially means a user uses posts in social media to display content via a timeline in a sequential time format. However, many times some events would be shared in different chronologies. We can therefore say, The social media narrative is a lengthy tale that has been divided up into postings for social

media platforms including Facebook, Twitter, LinkedIn, and Instagram. The postings are brief, but when combined, they might form a broader narrative with a theme. Therefore our questions are as follows:

- To what extent do social media posts make it possible to extract such features for a narrative explanation of a series of events?
- How do we explore the modeling narrative elements in social media in order to analyze user posts and understand his narrative?
- Can we model the social media posts via clustering and their intentions as a narrative element in an interactive generation framework?
- How to design a framework for interactively incorporating a user's social media posts to generate a narrative automatically?

We address these questions by creating new tasks that specifically model narrative elements in Social Media. We emphasize that what we care about is tweets clustering will be advantageous to identifying depression users and explaining the cause of depression. Our solutions to these challenges lead to a new framework known as **NarrationDep** (**Narra**tive Detec**tion** for **Dep**ression). The model may give improved results for determining the degree of sadness among Twitter users and infer explanations from the content of user messages. It can also construct new deterministic feature representations from training data. Our approach, which is based on explainable depression detection, may gather information about an individual's explainability from their tweets. The user's cluster tweets from the attention map are returned with explainable scores; the more significant the score, the more probable the cluster was significant and explained the narrative that contributed to depression classification.

1.4.4 Community Depression Dynamics Detection During COVID-19

Twitter and other online social media have seen increasing use throughout the shutdown. However, there is little effort put into identifying depression patterns at the state level or even more locally, like the level of a suburb. When necessary, such a detailed analysis of depression dynamics can assist authorities, including governmental agencies, in taking the appropriate steps more objectively in certain places. It also enables users to observe the evolution of depression over time to understand the efficacy of government programs

or the detrimental consequences of significant events. Accordingly, in this study we are looking for answers to the following questions:

- How does COVID-19 affect people's depression at the state-level temporal dimension?
- How does COVID-19 affect people's depression in local government regions in terms of time?
- Can we tell if the government's pandemic-era policies and actions had any impact on depression?
- Can we identify how major events throughout the epidemic affect depression?
- How well does the model capture the dynamics of depression in individuals?

With other researchers, we investigate community depression dynamics brought on by the COVID-19 epidemic in Australia in accordance with the difficulties. And my contribution and involvement centered around developing a depression classification model to satisfy one of our research goals of this thesis, to analyze the characteristics of depression, especially at the community level.; therefore, a novel method based on the term frequency-inverse document frequency (TF-IDF) and multi-modal data from tweets is presented. The goal of multi-modal features is to identify emotional, topical, and domain-specific depression. The TF-IDF is easy to use, scalable, and particularly successful in modeling a variety of lexical datasets, both large and small. Following the design of the depression model, Twitter data from the Australian state of New South Wales is gathered and fed into the model to extract the depression polarity that COVID-19 and associated events might impact during the COVID-19 timeframe.

1.4.5 Depression Detection at User-Level and Its Impact During Pandemic

In the midst of the COVID-19 epidemic, social media usage has increased as more people rely on acquiring the newest concerning COVID-19 information (136). Additionally, social media, including a variety of internet services and platforms like Facebook and Twitter, allow users to converse, interact and different information sharing. While social media can help distribute information, which could be beneficial in the fight against the epidemic, it has also been related to anxiety and depression (118). We hypothesize that

the COVID-19 pandemic and its social restrictions may affect depressed users and thus may be reflected in their daily tweets. Therefore in this study, we are aiming to address the following questions:

- How to distinguish between depressed and non-depressed users' tweets before and after the COVID-19 epidemic started?
- How effectively is the model detecting people,Äôs depression at the user level during the pandemic?
- How to extract various features from the same input and boost our model performance to identify depressed Twitter users?
- How does COVID-19 impact people's depression?

According to the research question, we propose a new hierarchical convolutional neural network (HCN) model to better model user tweet classification through social media. Our model can extract meaningful feature representations from the word and tweet levels. Eventually, we studied our model performance using a multi-channel convolution neural network CNN and multilayer perceptron MLP in word-level encoding to combine the advantages of two traditional neural network models. We found our model performance increased when we used two channels for word encoding, and we call this model with two channels HCN+.

1.5 Thesis Organization

The rest of the thesis is presented as follows:

Chapter 2 reviews the related work for depression detection, summarised some closely related work, including our previous works. And, understanding depression on online social networks could be carried out using Post-level behavioral analysis and User-level behavioural analysis.

Chapter 3 presents an approach for depression detection utilizes user post history summaries to choose the most relevant user-generated content automatically from online social media data.

Chapter 4 utilizes multi-aspect features from the diverse social media activity and tweets of the depressed user, In order to develop an explainable deep learning-based method for depression identification. The explainable model aids comprehension and offers suggestions for enhancing a deep learning model.

Chapter 5 presents, for the first time, an approach to modeling narrative elements to identify depression and understand crucial narrative elements for a user in social media and how they evolve and extract a story from a user's tweets.

Chapter 6 investigates in depth the mechanisms of community depression brought on by the COVID-19 epidemic in the Australian state of NSW. To identify depressive polarity from the text of tweets, a novel depression classification model based on multi-modal features and TF-IDF was developed.

Chapter 7 reports an investigation of depressed and non-depressed users' tweets during eight months before and after the start of the COVID-19 pandemic. A user classification model is proposed to automatically detect depressed users based on a hierarchical convolution neural network (HCN).

Chapter 8 summarises the contributions and suggests some potential research directions for future research.

CHAPTER 2

RELATED WORK

This chapter will explore related works that study mental illness in social media. We also emphasize how our work differs from these current methods. As a result, this chapter has primarily been conducted from 3 perspectives: 2.1 Social Media Dataset; 2.2 Depression Feature Extraction on Social Media; 2.3 Material and Methods For Online Depression Detection.

2.1 Social Media Dataset

Nowadays, most social media networks use textual data (posts) as the main media of communication. These social media networks have attracted the attention of the globe. People started using them to share, write and post about their daily activities and events. Social media network data is a valuable resource for researchers to study user behaviors (i.e., mental illness, violent content, political views, etc.). Social media platforms use different forms of textual data, i.e., Twitter uses tweets, Facebook uses posts, and Tumbler uses blogs, etc.

To collect data from social media networks, mostly the researchers have used two methods, (1) either using existing data sets proposed by the other researchers or (2) crawling data from social media networks. The former could be frustrating because of the limitation of the website's APIs, time-consuming issues and reliable ground truth.

Some researchers crawl data randomly from social media platforms using a simple vocabulary-based approach and developing and validating the terms used (vocabulary)

by users with mental illness. According to the work of (39), the author collected datasets from Twitter, and they crawled tweets containing keywords related to depression. They collected the tweets randomly that included negative and positive tweets. The same method approach is used in (164), by crawling tweets from Twitter, using phrases related to the antisocial disorder. Moreover, (51) have collected data from LiveJournal and have searched the LiveJournal community for depressed and non-depressed people. The depressed people data is extracted based on search communities by interest using words related to depression. While for non-depressed data, the authors explored based on words unrelated to depression.

However, One text post could not provide the truth of the actual mental situation of the users; because of that, some researchers rely on user self-report in their profiles. The authors in (213) have crawled 23 million tweets from 45000 users who self-reported as "depressed" in their profile to validate their hypothesis that they mention, "depressed user talked about symptoms on Twitter."

One of the limited data collected was by (85), which detected anxious and depressed users online from Twitter; They collected only 100 samples of users. They collected their data set by selecting the first 100 users who follow the Microsoft India student page on Twitter. Afterward, they extracted each user's tweets posted within 30 days; this approach suffers from less amount of data. On the other hand, a study conducted by (156) is one of the fascinating works that have gathered much ground truth data on depression. (1) A data collection of users who have been labeled as depressed. (2) A non-depressive data collection, where each user is labeled as such, (3) Depression-candidate data set, where the tweet was collected if it contained the phrase "depress," was gathered. The authors collected over 35 million tweets and 36,993 individuals who could be depressed.

2.2 Depression Feature Extraction on Social Media

Data in social media networks relates to all user contents, insights and information collected from individuals' behaviors in social media. This information reveals how people communicate with their connections. Data in social media is information that users share with the public, including user language and other metadata such as age, location, .etc. Generally, social data analysis involves two processes in the reviewed literature. The first step is collecting user-generated data on networking sites, and the second is analyzing that data. Exploring the social data typically takes place in real-time, used to assess variables, including influence, reach, and relevancy. In order to analyze

the social network data set, we have determined user behaviors to understand which feature extractors techniques we could use. Researchers who study online user behavior must consider many factors, such as how to distinguish between sentiment and data, time relevance (what is relevant today may not be relevant tomorrow), quality (how important particular messages and comments are to particular people), and how viral activity starts and spreads in social media websites. Generally, user behaviors on social media are various and diverse.

2.2.1 Social Network Post Features

Post-level features are the post's textual features that are gathered and converted into statistics. The post's linguistic content is described by these features as discussed in (36), and (67). For instance, the statements "I feel a bit depressed" and "I feel very depressed" utilize the term "depressed" yet convey two different emotions. The writers also looked at how people interacted with the posts, such as on Twitter (likes, retweets, and comments). Some studies examined post-level behaviors to predict mental issues by looking at tweets on Twitter to identify language associated with depression. A model has been created by (144) to find meaningful and relevant latent structure in a tweet. Similar to that, (156), X tracked several depressive symptoms indicated in a user's tweet. Moreover, analyzing users' posts from two platforms, Weibo and Twitter, conducted by (157). The authors used linguistic features and a sentiment Analysis using a Chinese psychological analysis program called TextMind. Another interesting post-level behavior studies done by (156) on Twitter by looking up words related to depression and antidepressants. (139) employed post-level behavior to identify anorexia; they examined terminology from the domain, including terms like an eating problem, anorexia, exercises, and food.

2.2.2 User Features

User-level Features are essential to look for in social media since they represent users' general behavior over several postings. Post-level features are obtained from a single post, whereas user-level features are obtained from many tweets at various times (178). The number of tweets, retweets, and/or user interactions with others is also used to obtain Twitter's user social engagement data. Generally, linguistic style in postings may be used to extract features (67) (213). From the collected data set, (156) defined six PTSD feature groups to describe each user thoroughly. The authors employed the number of tweets and social engagement as social network features. They have used user-shared

private information from social networks for user profile features. They showed that analyzing user behavior might help identify eating disorders. Similar to the previous work, (194) extracted user engagement and activities features in social media. Moreover, 70 features were extracted from two separate social networks by (157). (Twitter and Weibo). They took information from user profiles, posting histories, and user engagement data like follower and follower counts.

2.3 Material and Methods For Online Depression Detection

In the present era, a massive amount of unstructured data in the form of text is available on different platforms. E-mails, web pages, surveys, mobile apps, social media, and many more are the best examples. Text is a rich source of information, but usually, It is not easy and time-consuming to gain knowledge from it because it exists in unstructured and raw text form. In order to make this text data useful, it is necessary to normalize the data and extract valuable and related knowledge from it. Various methods have been used in literature to normalize data and extract features from it. Usually, raw text data is available in high dimensions, which need to be reduced using some feature reduction techniques such as Principle Component Analysis (PCA), Tabu search etc. Alternatively, instead of using raw text, we can create embedded words, which are vector representations of the words in the text. This process, known as word embedding, is often performed as a first step before training a machine learning model. One common method for creating word embeddings is Word2Vec, which utilizes shallow neural networks to learn the representations. (106). After the data has been transformed into a set of feature vectors, it can be used for predicting a class, target or label. Such predictions are also known as classification. Classification is predictive modeling that approximates the mapping function (f) between input variables (x) and discrete output variables:

$$(Y) => Y = f(x) \tag{2.1}$$

One of the classification methods belongs to supervised learning, where the targets also include the input data. In the case of text classification using supervised Machine Learning (ML) methods, the text is categorized into predefined categories (167). Text classification is one of the most commonly used ML tasks that is used in various business applications. Such "Text Classification " is an example of a supervised ML method

because, in such applications, a labeled data set of text documents is used to train a classifier. Rather than manually constructed rules, text classification by machine learning is achieved based on past observations. A machine learning algorithm can learn the relationships between different pieces of text by using pre-labeled examples as training data. Once the algorithm has learned these associations, it can then make predictions about the output for a new input. Some popular text classification fields are the emotions and sentiment analysis of textual data on the internet. With the help of text classification, one can easily understand audience sentiments and detect people's mental health using social media data. Text Classification is a very active research area both in academia and industry. In this section, we will cover some of the ML algorithms for text classification. The most popular ML algorithms that have been used for text classification in literature papers.

2.3.1 Traditional Machine Learning Algorithms For Depression Detection

Support Vector Machine: Although it is challenging to use computational linguistics techniques to replace in-person mental illness diagnosis completely, their successful use in tracking patients' progress and levels of depression during online therapy may give clinicians more knowledge and enable them to use interventions more skillfully and effectively. For instance, many researchers have used a supervised machine learning algorithm, the Support Vector Machines (SVM), as a classification method to predict the Mental State of the public. (36) was one of the first studies to detect depression on social media. Their study focused on finding the potential of social media to detect mental illness. Their research has shown the ability to use Twitter as a tool for individuals to assess and predict major depression. For that, they developed an SVMs-based classifier that could predict an individual's probability of depression before the reported onset of depression. (178) is another research in which the authors used SVMs as a model. Their purpose was to detect depression from the users' activities using SVMs as a classification method. SVMs are one of the most used classification methods in mental illness detection. SVMs have also been found to be a popular algorithm for Detecting Eating Disorders (EDs) in people. For example, (194) presented a study to detect EDs in people. Their study aimed to detect and characterize ED-ed communities on social media and classify ED-ed and non-ED-ed users. They compared many different parametric classifiers to determine the optimal classification algorithm. They found that SVMs give the best

classification performance. One psychological illness that has been diffused among many people and has devastating physical consequences is anorexia nervosa. (139) presented an approach based on ML algorithms that process social media users' texts to detect anorexia nervosa. To detect cases of anorexia and to evaluate their model, they explore four ML methods Logistic Regression, Random Forest, Support Vector Machine (SVM),), and Multilayer Perceptron (MLP). They found that SVMs outperformed the other classification methods. SVMs have always been the first option for many researchers to use as a classifier when they need to learn about data. Besides the fact that SVMs work well with unstructured and semi-structured data, another advantage of using an SVMs classifier is its regularisation capability. With the help of the regularization parameter, users can avoid over-fitting the model. With the bundle of advantages mentioned above, still, like other ML algorithms, SVMs suffer from limitations and weaknesses. For example, SVMs are not the right choice for large-scale data. In such cases, it will take longer to train. It also faces difficulty in understanding and interpreting the final model, variable weights and individual impact. (51) attempted to classify social media posts and online communities as either depressive or non-depressive through the use of Support Vector Machines (SVMs); they used Random Forest(RF) and SVMs. Their study found that RF is a powerful algorithm for multi-class classification that outperformed SVMs. (39) compared SVMs with Multiple Naive Bayes(MNB) for emotion detection. In their study, MNB outperformed SVMs. One of their observations was that supervised learning classification suffers from a limitation that cannot be accurate in predicting depression using text data.

Logistics Regression: Additionally, a statistical machine-learning technique has been adopted to study mental illness online, such as Logistics Regression(LR), which is widely used for various classification applications. One of the LR classification applications is text classification. LR has been used as a predictive analysis algorithm based on the probability concept. LR is used when the dependent variable(target) is categorical, as we mentioned before. For example, (156) proposed a new approach named the Multimodal Depressive Dictionary Learning Model (MDDLM) for identifying depressed users on Twitter. Using a large dataset of Twitter users, they examined the different online behaviors between depressed and non-depressed individuals to develop the model. Their model used Multi Dictionary Learning (MDL) combined with LR as a classifier. Another exciting research that used LR was by (67). The model was constructed using linguistic and behavioral characteristics, and the efficiency of the models was evaluated using varying observation periods. To train their model, they build a classification model using

LR. They compared the accuracy of the LR classification with different models based on different observation window times. LR has also been adopted to detect other mental illnesses, such as cognitive distortions. (163) tried to detect cognitive distortions using users' blogs on Tumbler social network. They used LIWS to extract textual features from the users' blogs. They experimented with LR with the other Machine Learning classifiers to detect cognitive distortion from the user profile on Tumbler. The best result was achieved when they used RELIEF (80) with Logistic Regression.

Stress could be a challenging mental illness to detect through social media, but (91) did exciting work. They presented a framework to detect the weakly information and interaction of a user on social media in order to detect stress. LR and the other classification methods showed excellent performance in their research. LR is not a good choice for non-linear problems; in such cases, a more complex model can outperform it. For example, (139) compared different classifiers for the text analysis written by a user on social media and detects anorexia nervosa. They explore four methods: LR, SVMs and two other Machine learning methods. They found that, as a result, LR falls behind SVMs.

Naive Bayes: Naive Bayes(NB) has also been a popular choice for researchers to solve different classification problems. (39) compared two ML algorithms as classifiers methods: Multinomial Naive Bayes (MNB) and SVMs for text-based emotion AI that applied to detect depression in Twitter. They found that MNB outperformed SVM. (Akshi Kumar et al.). On the other hand, (156) used their model to identify sad Twitter users and compared the performance of their model and the baselines in terms of the four selected measures; one of these methods was NB. Results show that NB and the other classification methods failed their new model. (164) in their work to detect antisocial behavior on social media showed that NB was giving less accuracy as a classifier compared to SVMs and RF. (163) also showed that NB failed against LR in order to detect cognitive distortions.

TF-IDF A numerical statistic called term frequency-inverse document frequency (TF-IDF) or (TFIDF) is meant to indicate the significance of a word in a list or corpus. TF-IDF is simple to calculate and contains several fundamental metrics for identifying the most descriptive term inside a text. It makes it simple for TF-IDF to determine how similar the two texts are. These advantages attract many researchers to use TF-IDF to build their models. (36) adopted TF-IDF in their work. They selected the top words with high TF-IDF values after calculating the TF-IDF for these words in Wikipedia to exclude a general term. (164) Used (TF-IDF) in his model to see the difference of the

result combining (TF-IDF) with several machine learning models. They were able to achieve high precision and accuracy with five different Machine learning algorithms. Still yet, TF-IDF is suffering from some limitations. Since it is based on the bag-of-words (BoW) paradigm, it is unable to account for factors like a text's location, semantics, co-occurrences across documents, etc. Due to this, TF-IDF is only effective as a lexical-level feature and is unable to collect semantic information (e.g., as compared to topic models and word embedding).

Latent Dirichlet Allocation (LDA) : Obtaining topics from a text brought a researcher's attention, and the most often used topic modeling technique is Latent Dirichlet Allocation (LDA). LDA assumes that a mixture of topics creates documents. Then these topics generate words based on their distribution of probability.

Many researchers adopted LDA in their approach to detecting mental illness, and it is found to be very successful in depression detection according to (156). The authors used Latent Dirichlet Allocation (LDA) model to extract the topic of document distribution. While the authors in (178) used LDA to extract topics from each user's tweets. They found it useful to evaluate the degree of depression. One of the reviewed studies that used LDA for mental illness detection was carried out by (144); the authors investigated the possible utility of more advanced topic modeling in automated depression detection. They explore the use of supervised topic models in linguistic signal analysis to diagnose depression, providing promising results across multiple models. They show that LDA can disclose meaningful and potentially useful latent structures. The authors in (139) used LDA in their proposed approach to detect anorexia nervosa. In the result, they found out that the feature based on linguistic inquiry and word count (LIWC) dictionary and LDA feature improved the result of their model. These traditional topic modeling, however, experience a significant loss of performance over short texts because of the lack of word co-occurrence information in each short text. According to the reviewed literature, all the previous studies for detecting mental health issues through social media used traditional topic modeling for short text. Therefore, the short text may affect the performance of traditional topic modeling. Short text, such as tweets on Twitter and posts on Facebook, have less word co-occurrence information. Because the fewer words in each text, most of each text generates one topic. Words' statistical information between texts cannot capture words related in semantics but rarely co-occur. With these issues of the short text, many studies proposed the traditional topic modeling for short text, and none of these new models have been adopted in the reviewed literature to detect mental illness through social media.

Hybrid model: (85) Proposed a new model to predict anxious depression disorder in real-time tweets based on the linguistic and user's post behavior. For classification, they proposed a new classifier that combines the results of three ML classifiers MNB, Gradient Boosting (GB) and RF. The combined classifier was given the name of the ensemble vote classifier. (91) presented a framework to detect the weakly information and interaction of a user on social media. They built a hybrid model using a factor graph with CNN to understand users' posts content and interactions. They compared their model FGM+CNN (FGM) with three comparison methods, i.e., LR, SVMs, RF and Gradient Boosted Decision Tree(GBDT). All the methods showed a good performance, but their model FGM outperformed them.

2.3.2 Artificial Neural Networks for Depression Detection

One of the massively parallel systems with a large number of simple interconnected processors is Artificial Neural Networks (ANN). It is a software implementation of the neuronal structure of our brains. The brain contains neurons, which are human switches. They may change their output state depending on the strength of the neuron's electrical or chemical input. The neural network in a human brain is a highly interconnected neuron network, where the output of any single neuron can be the input of thousands of 'other neurons. ANN is simplifying the behaviors of the human brain and is trying to mimic them. The ANNs can be trained either in a supervised manner or an unsupervised manner. In a supervised method, matched samples of input and output information are provided for training ANNs. Desired output from the trained ANN is then obtained for a given input (167). This method is referred to as a network because it can be visualized as multiple layers of interconnected nodes. The first layer, known as the input layer, comprises nodes representing all the input features. The final layer, called the output layer, comprises one or more nodes that predict the class. There exist one or more 'hidden' layers of the node between the input and the output layers. During the tuning process of training ANNs, the number of nodes in the hidden layer is determined. In ANNs, artificial neurons are stimulated by an activation function. This activation function must have a switch on the characteristic in identification tasks (e.g., detecting health problems). In other words, the output must change the status once the input is greater than a certain value. One of the most used activation functions is the sigmoid function:

$$f(x) = \frac{1}{1+e^{-x}} \tag{2.2}$$

ANNs have the ability to work as an ML algorithm for making different decisions. Because of this capability, many researchers have adopted their models in various research areas of the ML field, such as mental illness detection. ANNs contain multiple different architectures that have been used as classification methods. Each one has his own principles and rules.

LSTM: Stands for Long Short-Term Memory in ML algorithms. LSTM is a type of Recurrent Neural Network (RAN) that is one of the architectures of ANNs. It has short-term memory, which is why it uses previous persistent information in the current neural network. LSTM is networking with loops in them, which is a model that extends the memory of RNN. For anything that has a sequence, LSTM is an excellent tool because the meaning of a word depends on the preceding ones. The power of LSTM attracts researchers to use it in mental illness detection. For imbalanced data that is extracted from social media, (30) proposed a model to detect depression using Bidirectional LSTM, which is just putting two independent LSTMs together.

Multilayer Perceptron:A perceptron in ANN is a simple model of biological neurons. Perceptron is also the name of an early algorithm for the binary classifier based on supervised learning. The perceptron algorithm has been developed to distinguish visual inputs, classify subjects into one of two types and divide groups with a line. Multilayer Perceptron (MLP) sometimes could not be the best choice for text classification. It has been compared with other classifiers to detect mental problems. The result showed that Multilayer Perceptron in the classification accuracy (163) (139). In addition, deep learning has gained significant attention in recent years for its applications in the healthcare sector. For example, deep learning in medical image processing and diagnosis has led to impressive results in extracting valuable and actionable information from complex and diverse data sets. Deep learning has the benefit of excellent iterative learning, and automatic latent representation optimization from multi-layer network structures (125). This encourages us to combine social media users' rich and diverse behavioral patterns with superior neural network learning capability. Additionally, it was suggested by (90) to identify Twitter users based on depression using deep learning techniques with pre-trained language representation models, including transformers models, explored for depression detection using the Instagram platform. Their methodology uses CNN-based deep learning with the Bidirectional Encoder Representations from Transformers (BERT) model to use textual and visual data from posts and pictures. Also, (196) intended to

investigate the capability of deep-learning techniques to identify depression using data from Chinese microblogs. They discovered that language representation models that had already been trained, like BERT, performed the best.

2.4 Summary

This chapter first explained traditional depression detection and how deep learning techniques usually depend on large datasets with reliable co-occurrence statistics, which It may not be found in short and sparse texts. Due to potential conflicts between effectiveness and explainability, most current models need to explain prediction. The explainable model can provide insight into how to enhance and assist understanding in a deep learning model. Moreover, previous works of literature focused on studying user behavior rather than taking cues from user-generated content, such as the text they share, making it highly challenging to do well in classification. Additionally, these models do not effectively identify depressed users at the user level, so they are prone to incorrect predictions. Recent research suggests the critical challenges for identifying depression online are:

- The diversity in the user's behaviors on social media.
- Posts diversity per user.
- Explainable for automated depression Detection.
- Study depression dynamics at the state level due to COVID-19.
- Detect depression at the user level due to COVID-19.

These research gaps are the primary motivation for this thesis and related online depression detection research. Moreover, this thesis aims to develop novel techniques for addressing the challenges of online depression detection. To overcome diversity in the user's behaviors and post diversity per user on social media challenges, we have proposed a new computational approach for automatically detecting depression in chapter 3. Using a hybrid extractive and abstractive summary process on the order of all user tweets, the new approach first chooses relevant content, producing more concise and important content. We have also utilized multi-aspect features from the diverse social media activity and tweets of the depressed user, In order to develop an explainable deep learning-based method for depression identification in chapter 4. The explainable model

aids comprehension and offers suggestions for enhancing a deep learning model. In chapter 5, we will dive deeper for explainability in depression detection and will present, for the first time, an approach to modeling narrative elements to identify depression and understand crucial narrative elements for a user in social media and how they evolve and extract a story from a user's tweets. The challenge in studying depression dynamics at the state level due to COVID-19 is tackled by proposing a novel technique built on (TF-IDF) and multi-modal features from tweets in Chapter 6. Eventually, the problem of detecting depression at the user level due to COVID-19 is addressed in chapter 7 by developing a new approach based on Hierarchical Convolutional Neural Network (HCN). The new approach can detect depressed users occurring within the COVID-19 time frame.

CHAPTER 3

LEARNING MULTI-MODALITIES WITH USER POST SUMMARIZATION

3.1 Background and Motivation

Major depressive disorder (MDD), also known as depression, is among the most prevalent psychiatric disorders globally[1] which leads to a substantial economic burden to the government. Early detection of depression is crucial for effective policing and security agencies because it could adversely impact innocent citizens, e.g., mass shootings (111) whose cause has usually been attributed to mental health problems. As depression is a disorder that requires self-reporting of symptoms frequently, social media posts such as tweets from Twitter provide a valuable resource for automatic depression detection.

People with depression often tend to hide their symptoms to avoid revealing that they are afflicted (146) or they find it difficult to consult a qualified diagnostician (162). Many go to online social media to express their underlying problems. One of the reasons is that they are willing to share their problems within their friend network thinking that they might offer help or advice. Sometimes they also implicitly leave clues which could point towards a state of depression or onset of early depression. Manually finding such users online by scanning through their posts will be very time-consuming. Therefore, the challenge is how we can propose effective computational techniques to automatically

[1] https://www.who.int/news-room/fact-sheets/detail/depression

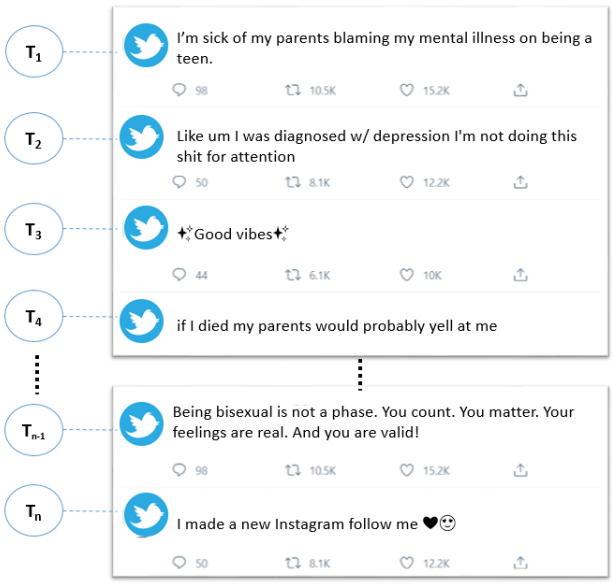


Figure 3.1: A sample of depressed user tweets.

find such users online. Depression detection on Twitter is a promising and challenging research problem. Building an approach that can effectively analyze tweets with self-assessed depression-related features can allow individuals and medical specialists to understand the depression-levels of users. Researchers have proposed unsupervised methods to automatically detect depression online (213; 178) using computational techniques. One of the shortcomings in existing methods is that they tend to use every social media post of a user. We argue that this is not necessary because it tends to make the automatic depression detection system inefficient and even degrade the performance, e.g., dealing with "curse-of-dimensionality" and that these irrelevant posts may have a dominating impact more than depression-sensitive content. It is common for every user to share a varied set of posts online not just depression-related and we depict this user pattern through an example in Figure 3.1 which shows a variety of posts which might not be even relevant to depression. As a result, we need methods which could help dampen the impact of content which might eventually not help the classifier. We also need an effective feature selection strategy so that we could detect patterns from users which are implicit/latent which, unfortunately, cannot be modelled well by adopting simple term-frequency mining or modeling the surface-level features.

In single document automatic text summarisation (172), the goal is to condense the document into a coherent content without losing any semantic information. Our proposed a novel computational framework trains on the content obtained after performing extractive-abstractive user-generated content summarization that helps select

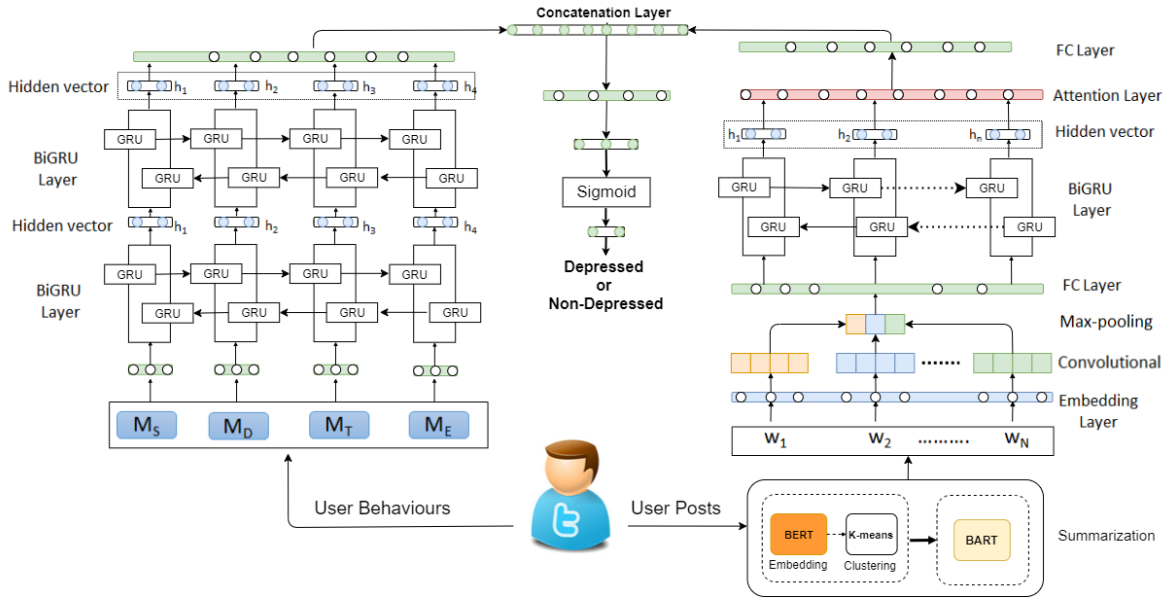


Figure 3.2: Proposed Framework (DepressionNet) for Depression Detection on Social Media

the most useful features for the classifier. The reason why we adopt the summarization approach is that it enables us to preserve the most salient content for every user and condenses them giving us a summary of the content. We use this summary in our novel deep learning framework which is based on Convolutional Neural Network (CNN) and Bidirectional Gated Recurrent Units (BiGRU) with attention. Exploiting the capability of the CNN network enables to model the features more faithfully. However, CNN usually performs suboptimally to capture the long term dependencies which are usually word order information. To mitigate this shortcoming, we introduce a bidirectional GRU model which belongs to the family of Recurrent Neural Network (RNN). We have used BiGRU to learn long-term bidirectional dependencies from backward and forward directions in our data because it has shown to perform well than a unidirectional GRU model. To further capture the user patterns, we have introduced various user behavioural features such as social network connections, emotions, depression domain-specific and user-specific latent topic information, and applied stacked BiGRU. A stacked architecture is obtained by having multiple BiGRUs runs for the same number of time steps which are connected in such a way that at each step the bottom BiGRU takes the external inputs. The higher BiGRU takes inputs externally as the input state output by the one below it. As pointed in (61) residual connections between states at different depth are also used to improve information flow. Finally, we combine both user behaviour and post summarization in

our framework, which we call DepressionNet, that consists of two shared hierarchical late fusion of user behaviour networks and a posting history-aware network. We conducted comprehensive experiments, and our model significantly outperforms existing approaches (+3% in Acc and +6.5% F-Score); thus, our **key contributions** are as follows:

1. We propose a novel deep learning framework (DepressionNet) for automatic depression detection by combining user behaviour and user post history or user activity.

2. We apply a abstractive-extractive automatic text summarization model based on the combination of BERT-BART that satisfies two major requirements:

    - wide coverage of depression relevant tweets by condensing a large set of tweets into a short conclusive description.
    - preserving the content which might be linked to depression.

3. To further make our prediction reliable, we have used information about user behaviour. To this end, we have developed a cascaded deep network that concatenates the behavioural features in different layers.

3.2 Preliminary on Learning Multi-Modalities with User Post Summarization

Social media is one the platforms which could help discover and later propose ways to diagnose major depressive disorders. Researchers studied the effects of social media to predict depression since these platforms provide an opportunity to analyze individual user and state of mind and thoughts (32; 36; 38). Digital records of people's social media behaviours, such as Facebook and Twitter, can measure and predict risks for different mental health problems, such as depression and anxiety. Additionally, these websites have been shown to allow machine learning and deep learning algorithms to be developed.

3.2.1 Early Depression Detection

In the context of online social media, various studies in the literature have showcased that analysing user content and user textual information in social media has helped achieve some success for depression detection and mental illness. Lately, shared tasks

such as CLEF eRisk, Losada et al. (97; 98) have included automatically detecting depression as early as possible from a users' posts. They discovered that there are features that capture differences between normal and depressed users. Several studies aim to analyze emotion, user network, user interactions, user language style and online user activities as features to identify depression on social media (212; 179; 156; 158). However, all these features are treated as an individual measurable property in different machine learning algorithms. For instance, to detect depressed users online (36) used the records of user activities with support vector machine (SVM) model for classifications and found the possibility to recognize depression symptoms through examining user activities on Twitter. In (114) authors tried four different classifiers to classify user activity and found naive Bayes' model performed well. In (178) authors also used user activities online; however, they determined that a detailed evaluation is needed to estimate the degree of depression through user activity history on social media.

One of the important mental health issues for new mothers is postpartum depression (PPD) (110; 38; 155). According to (155), this depression could affect fathers too. The authors collected social media posts from fathers. They used the SVM model as a classifier and found that fathers could at risk of PPD. Other researchers study some other symptoms of depression, such as suicidal thoughts at an early stage. The authors in (5) studied that people with suicidal thoughts leave a note on social media; therefore, they applied a classifier on posts, and the title on Reddit to differentiate suicidal and non-suicidal notes. They used term-frequency and inverse-document frequency (TF-IDF) as features and they found that logistic regression to be the ideal classifier for detecting the suicidal posts online.

Recently, some studies have started to target depressed user online, extracting features representing user behaviours and classifying these features into different groups, such as the number of posts, posting time distribution, and number followers and followee. In (133), they extracted different features and classified them into three groups, user profile, user behaviour and user text and used multi-kernel SVM for classification. In (156) the authors proposed a multi-modalities depressive dictionary learning method to detect depressed users on Twitter. They extracted features from a depressed user and grouped these features into multiple modalities and proposed a new model based on dictionary learning that is capable of dealing with the sparse or multi-faceted user behaviour on social media.

The above-mentioned works have some limitations. They mainly focused on studying user behaviour than taking cues from user generated content such as the text they share

which make it extremely difficult to achieve high performance in classification. These models also cannot work well to detect depressed user at user-level, and as a result, they are prone to incorrect prediction. Our novel approach combines user behaviour with user history posts. Besides, our strategy to select salient content using automatic summarization helps our model only focus on the most important information.

3.2.2 Deep Learning for Depression Detection

Use of a deep neural network (DNN) such as convolutional neural networks (CNNs) and long-short short-term memory (LSTMs) (62) have made notable progress in detecting mental illness on social media (168; 182; 24; 140). Deep learning has obtained impressive results in natural language processing (NLP) tasks such as text classification and sentiment analysis. Several works in the literature have concentrated on analysing user content and user textual information via deep learning models. For instance,Shen et al. (158) explore a challenging problem of detecting depression from two online social media platforms, Weibo and Twitter; therefore, they introduced the cross-domain DNN model using adaptive transformation & combination features. Their model can address the heterogeneous spaces in various domains comprising of several features. To incorporate the users' behaviour, they extracted various features from both domains and classified them into four groups. Recently, different text classification models based on deep learning have been developed to determine if a single tweet has a depressive propensity.

In (177) the authors targeted at the early detection of anorexia and depression on eRisk 2017 dataset, and in (176), the authors find that CNN outperforms LSTM. They used the CNN model with GloVe (134) and fastText (16) where for each record, they vectorized only the first hundred words. They have also trained other models using a broad range of features such as linguistic metadata and LIWS, and they found that linguistic metadata and word embeddings preformed well. This work inspired other researchers to classify posts on online forums to identify depression-related suicides (211). For this task, they proposed a large-scale labelled dataset containing more than 116,000 users and proposed a model based on CNN architecture to identify depressed users concentrating on learning representations of user posts. Due to a large number of posts of each user, they created posts selection strategy to select which posts they used to train their model (earlier, latest and random posts), and among these three selection strategies, selection random posts gave better performance. On the other hand, RNN showed potential when it is applied to identify depression online. To classify depressed and healthy speech among patients screened for depression, Al Hanai et al. (4) utilize an

RNN model. They combined LSTM based model to concatenate the sequence and audio interviews of the patients. Gui et al. (58) fused two different data inputs, instead of text and audio. They observed that with the consideration of visual information of a post on Twitter, the meaning of that post is easy to determine. Shen et al. (156) proposed a model to detect depression using fusing images and the posts of a user in social media. To compute continuous representations of user sentences, they used Gated Recurrent Unit (GRU), and for images, they used 16-layer pre-trained Visual Geometry Group (VGGNet).

3.2.3 Automatic Text Summarization

Sequence-2-sequence models have been successfully applied to many tasks in NLP including text summarization (115; 94; 171). Recently, transformer (184) model has become popular for automatically summarising text documents. Liu et al. (95) proposed a model that can hierarchically encode multiple input documents and learn latent relationships across them. Recently, pre-trained language models have been commonly integrated into neural network models, such as BERT (41) and BART (88). For tasks in NLP such as text summarization, these trained models have achieved state-of-the-art performance. In essence, the BERT model is based on an encoder-decoder network transformer. BERT has been adopted in the medical field as an extractive summarization approach, to summarise patients medical history (189). Recently the other pre-trained language model BART has attracted many researchers since it obtained new state-of-the-art performance in summarization task. BART contains a two-part bidirectional encoder and auto-regressive decoder. The robust of BART model lead (59) construct a corpus for text summarization in the Russian language, and they utilize a pre-trained BRT model for Russian language summarization. And their BART works extremely well, even though it was not originally intended for Russian language text summarization.

In this chapter, we propose a novel depression detection model called (DepressionNet) using online user behaviours and summarization of his posting history. Our motivation for applying automatic text summarisation mainly comes from the fact that summarisation can aid our model to primarily focus on those content that is condensed and salient. To the best of our knowledge, we are the first ones to use automatic text summarization for depression detection on online social media.

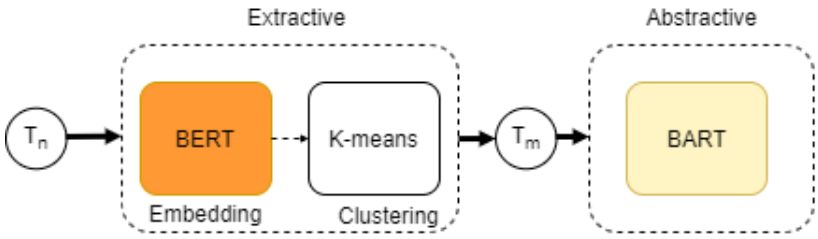


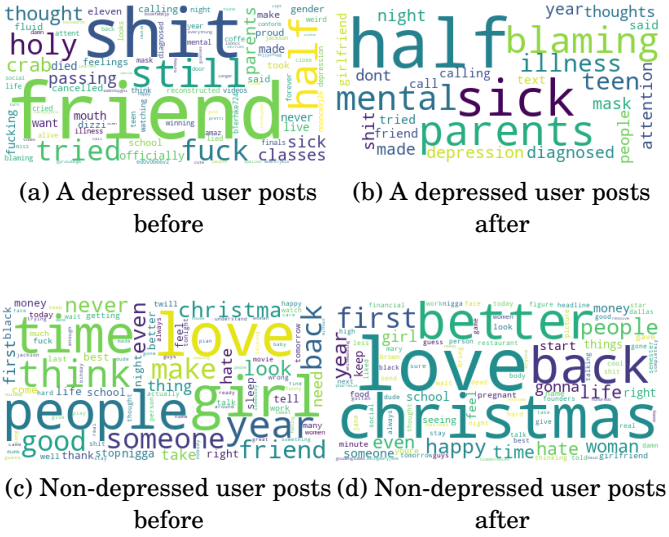
Figure 3.3: Diagram to illustrate the process of user posts summarization.

3.3 Our Novel DepressionNet Model

In this section, we present our proposed novel framework, DepressionNet for the task of automatic depression detection by fusing the user behaviour and the user post history. Figure 3.2 depicts the full model components. User posts can be abundant, redundant and may contain irrelevant information that might not give useful information to the computational model. This poses a significant challenge to effectively employ the knowledge learned from a user on social media which is already characterised by short and noisy content. Consider that a user U_i has posts $[T_1, T_2, ..., T_n]$ from the user activity history, where the total number of posts is n. Each T_i is the i^{th} user-generated content. Our goal is to assign a label $y_i \in \{depressed, non\ depressed\}$ to the user U_i signifying a binary assignment whether the user is depressed or not depressed. To realise our goal, we have fused the user behaviour and user post history of each user U_i. The abstractive-extractive summarization can be defined as extracting and summarizing the user generated content T_m from user history of T_n tweets such that $T_m \leq T_n$. Besides, we incorporate user behavioural information social network, emotions, depression domain-specific and topic modelling denoted as M_S, M_E, M_D, and M_T, respectively.

3.3.1 Extractive-Abstractive Summarization

User post history plays a vital role in observing the progression of depression, thus, we considered the problem of depression detection by analyzing depressed user posts history to better understand the user behaviour for depression. Single-user post history summarization is the task of automatically generating a shorter representation of user historical post while retaining the semantic information. Figure 3.4 depicts the importance of text summarisation on a real user-generated content. We notice that a user who is depressed, the module distills the redundant and non-informative content, e.g., “still”, “eleven” have been removed after summarisation from a depressed user and salient words such as “sick”, “mental” have become prominent. We also observe the same pattern



(a) A depressed user posts before

(b) A depressed user posts after

(c) Non-depressed user posts before

(d) Non-depressed user posts after

Figure 3.4: A word cloud depicting words from depressed and non-depressed users before and after extractive summarization. We show qualitatively that summarisation helps in selecting the most salient or focused content.

for non-depressed user where the focus only remains on most non-redundant patterns after summarisation. We argue that adopting summarisation technique is useful for the model to remove many irrelevant content so that we could focus on the most important information associated with a user.

We present a framework that incorporates an interplay between the abstractive and extractive summarisation. The reason why our model is based on extractive and abstractive framework is that extractive model helps automatically select the user generated content by removing redundant information. Abstractive framework further condenses the content while preserving the semantic content. Given that we might have a large amount of tweets associated with each user, this approach mainly helps reduce the amount of redundant information and noise in data. Our model relies on BERT-BART summarisation model which is a strong automatic text summarisation model based on contextual language models, which are very popular models for modeling text.

On the user posts (T_n), we have applied extractive tweets summarisation to select the most important tweets (T_m) from full set of user posts. The unsupervised extractive summary takes a pre-trained BERT model to perform sentence embedding (109). Figure 3.3 depicts the design of our extractive-abstractive automatic summarization. As the BERT is trained as masked-language model, it gives us a vector representation ($[W_1, W_2...., W_j]_{T_i}$) for each tweet as the output which is grounded to tokens instead of

sentences. We then perform k-means clustering on high dimensional vector representing semantic centers of text, the cluster centres are selected to extract most important tweets T'_i. For example, user A posts n tweets $T = [T_1, T_2, T_3, ..., T_n]$, the extractive summarization will return m tweets $T' = [T_1, T_2, T_5...., T_{n-1}]$ whereas $T' \subseteq T$, i.e., in Figure 3.1, T_3 and T_n are depression irrelevant tweets, extractive summarization excludes tweets T_3 and T_n.

Once, we have summarised the posts (T'_m) for each user (U_i) using the extractive summarisation, we can further condense the remaining redundant information that might gone undetected during the extractive summarisation phase. To this end, we apply the BART model which perform abstractive text summarisation. BART is denoising sequence-to-sequence autoencoder that uses transformer structure. The BART structure consists of two components: an encoder and a decoder. The encoder component is a bi-directional encoder compliant with the BERT architecture and the decoder component is an auto-regression decoder that follows GPT settings. We have used BART-large model (88) which has originally been fine-tuned on CNN/DM dataset as abstractive summarization.

BART produces word embeddings that represents the summary of user posts at the word level. The word embeddings are then used as input to the stacked Convolutional Neural Network (CNN) and Bidirectional Gated Recurrent Units (BiGRU) with attention model to capture sequential information, such as the context of a sentence. The attention mechanism is advantageous in this scenario because the model helps focus on relevant words. The summary can be expressed as $S = \{w_1, w_2, .., w_N\}$, where $S \in \mathbb{R}^{\mathbf{V} \times N}$, N represents the summary length, V represents the size of the vocabulary. Each word w_i in S is transformed to a vector of word x_i using Skipgram model available in word2vec library. We have used the pre-trained with 300-dimensional[2] embeddings. The embedded summary sentence can be represented as:

$$X = \{x_1, x_2, .., x_N\} \tag{1}$$

A weighted matrix of word vector will be utilized as embedded layer output and is input to the convolutional neural network (CNN) followed with layer of max-pooling and ReLU. The goal of CNN is to extract the most relevant embedded summary sentence features. The word vector representation of a tweet is typically complex, thus, the word vector dimension is regularly taken by the CNN layer with kernel dimensions which

[2]https://github.com/mmihaltz/word2vec-GoogleNews-vectors

extract important features by learning the spatial structure in summarized text through pooling. Finally, we add a fully connected (FC) layer which serves consolidated features to the BiGru.

3.3.1.1 The Bidirectional Gated Recurrent Unit (BiGRU) Layer

The resulting features from CNN layer are passed to the BiGRU, which is a RNN that can capture sequential information and the long-term dependency of sentences. Only two gate functions are used which are reset and update gates. Update gate has been used to monitor the degree with which the previous moment's status information has transported into the current state. The higher the update gate value, the more the previous moment's status information is carried forward. The reset gate has been used to monitor the degree with which the previous moment's status information is overlooked. The smaller the reset gate value, the more neglected the context will be. Both the preceding and the following words influence the current word in the sequential textual data, so we use the BiGRU model to extract the contextual features. The BiGRU consists of a forward GRU and a backward GRU that are used, respectively, to process forward and backward data. The hidden states obtained by the forward GRU and the backward GRU for the x_t input at time t are forward(h_t) and backward(h_t), respectively.

$$\text{forward}(\ h_1) = GRU(x_t, \text{forward}(\ h_{t-1})) \tag{2}$$

$$\text{backward}(\ h_t) = GRU(x_t, \text{backward}(\ h_{t-1})) \tag{3}$$

The combination of the hidden state that is obtained from the forward GRU and the backward GRU $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ is represented as h_t as the hidden state output at time t, and the output of the i^{th} word is $h_t = ((\text{forward}(h_{t-1}) \oplus (\text{backward}(h_{t-1}))$.

The attention mechanism helps the model to assign different weights to each part of the input and to reflect the correlation between features and performance results. Let H be a matrix consisting of output vectors $[h_1, h_2, h_3, ..., h_N]$ which we obtain from BiGRU layer, where N here is the length of the summary sentence. The target attention weight u_t at timestamp t is calculated using the vectors h_t vector $u_t = \tanh(h_t)$.

We can get the attention distribution a_t, computed using a softmax function as $a_t = \frac{\exp(u_t)}{\sum_{t=1}^{m} \exp(u_t)}$. A user summarized posts attention vector $\bar{s}_i$ is calculated as the weighted sum of posts summarization features, using the dot product to sum products of a_t and h_t as follows: $\bar{s}_i = \sum_{t=1}^{m} a_t \bullet h_t$, where $\bar{s}_i$ is the learned features for the summary.

Table 3.1: Summary of User Behaviour features

Modality	Features Description
Social Network	1-Posting time distribution for each user. 2-Numeber of followers and friends (followee). 3-Number of tweets, re-tweets and tweets length.
Emotional	1-Number positive, negative or neutra emojies. 2-(valance, arousal, dominance) score for each tweet 3-Calculate first person singular and plural from each tweet.
Domain Specific	1-Count of depression symptoms occurring in each tweet 2-Count of antidepressant words occurring in each tweet
Topic	1-Using the LDA model

3.3.2 User Behaviour Modelling

We have also considered user behaviour features and grouped all features into four types which are user social network, emotions, depression domain-specific and topic features obtained using a probabilistic topic model, Latent Dirichlet Allocation (LDA). We have shown more details in Table 3.1. These four feature types are also used in (156). However, we have not considered *User Profile feature* and *Visual feature* types due to missing values. We have extracted these feature-types for each user that are described as below:

I- Social Network Features: We extracted several features related to user social interactions such as the number of followers and friends (followee). We have also considered user posting behaviour such as the number of tweets, retweets, and the length of tweets. Besides, we have extracted the posting time distributions as features, i.e., the number of tweets in an hour per day by user U_i.

II- Emotional Features: User emotion plays important role in depression detection. We have considered emotional features such as valence, arousal, and dominance (VAD). The lexicon includes a list of English words and their valence, arousal, and dominance scores as features (17). We create a dictionary with each word as a key and a tuple of its (valance, arousal, dominance) score as value. We then parse each tweet and calculate VAD score for each tweet using this dictionary. We then add the VAD scores of tweets for a user, to calculate the VAD score. Tweets are rich in emojis and carry information about the emotional state of the user. Emojis can be classified as positive, negative or neutral, and can be specified as Unicode characters. For each of positive, neutral, negative type,

first, we count their appearance in all tweets by user U_i.

III- Domain-Specific Features: The features which are often domain-dependent produce faithful classification results. We have considered two different types of domain-specific features which are depression symptoms and antidepressant related. For depression symptoms, we count the number of times any of the nine depression symptoms for DSM-IV criteria for a depression diagnosis (46) is mentioned in tweets by user U_i. The symptoms are specified in nine lists, each containing various synonyms for the particular symptom. For each depression symptom, we count the number of times a symptom S_j appeared tweets by user U_i. For antidepressant, we created a separate list of antidepressant medicine names from Wikipedia[3], and count how many times the antidepressant name is mentioned by user U_i.

IV- Topic related Features: Topic modelling uncovers the salient patterns (represented as distributions over the words in a tweet) by user U_i under the mixed-membership assumptions, i.e., each tweet may exhibit multiple patterns. The frequently occurring topics plays a significant role for depression detection. We first consider the corpus of entire tweets of all depressed users and split each tweet into a list of words followed by assembling all words in decreasing order of their frequency of occurrence. We have removed stop words from the data. We applied unsupervised Latent Dirichlet Allocation (LDA) (15) to extract the latent topic distribution. For each user U_i, we compute, how many times each of words occurs in user's tweets separately.

To obtain fine-grained information, we have applied stacked BiGRU for the each of multi-modal features. In our experiments, we have considered two BiGRU that capture the behavioural semantics in both directions backwards and forward for each user followed by fully connected layer as shown in Figure 3.2. Suppose the input which resembles a user behaviour be represented as U_i=[$m_1, m_2, m_3, ..., m_N$] for i^{th} user. The outcome of behaviour modelling is the high-level representation that captures the behavioural semantic information and plays critical role in depression diagnosis (see ablation study section 3.4.4).

3.3.3 Fusion of User Behaviour and Post History

Figure 3.2 shows that the overall network consists of two asymmetric parallel networks (user post history network and user behaviour network) that consists of two shared hierarchical late fusion networks and a posting history-aware network that is combined with fully connected (FC) layer. The hierarchical temporal-aware network coalesces

[3]https://github.com/hzogan/DepressionNet/blob/main/domain_specific

multiple fully connected layers to integrate user behavioural representation and user posting (history-aware posting temporal network). For example, for user U_i, we have extracted a compact feature representing both behaviour and user posting history followed by a late fusion. The resulting framework models a high-level representation that captures the behavioural semantics. Similarly, the user post history comprises of representations extracted from user history that represent the gradual growth of depression symptoms. We have concatenated both representations to generate a feature map that considers both user behaviour and reflection of user historical tweets. The output of the DepressionNet network is a response map that denotes the similarity score between the depressed user and non-depressed user. As the network coalesces multiple hierarchical fully connected convolutional layers, thus, the network may have a different spatial resolution. To overcome this challenge, we exploit the max-pooling to down-sample the shallow convolutional layer to the same resolution as the deep convolutional layer. The hierarchical integration of user behaviour network results in a significant improvement of performance (see ablation study Section 3.4.4).

3.4 Experiments and Results

In this section, we will describe our experimental setup in detail followed by comparisons with the state-of-the-art models.

3.4.1 Baseline Methods

We compare our proposed method with various strong comparative methods. Our comparative methods range from those that have been proposed for depression detection and for general text classification models because our setting also resembles that of binary text classification. For user behaviour features, we have used methods that have been applied for detection of mental illness. Multi-modal Dictionary Learning Model (MDL) has been proposed to detect depressed users on Twitter (156). They used a dictionary learning to extract latent data features and sparse representation of a user. Support Vector Machines (SVM), is a popular and a strong classifier that has been applied on a wide range of classification tasks (75) and it still remains a strong baseline. Naïve Bayes (NB) is a family of probabilistic algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between instances (40; 3). Given the popularity of contextual language models trained using modern deep learning methods, we have also investigated three popular pre-trained models, which are, BERT (41), RoBERTa (96)

Table 3.2: Summary of labelled data used to train depression model

Description	Depressed	Non-Depressed
Numer of users	2159	2049
Number of tweets	447856	1349447

and XLNet (209) for summarization sequence classification. We fine-tuned the models on our dataset. We also compare with GRU+VGG-Net with cooperative misoperation multi-agent (COMMA) (58), where the authors proposed a model to detect depressed user through user posts (text) and images. They constructed a new dataset that contains users tweets with the images, based on the tweet ids (156).

3.4.2 Dataset

We have used a large-scale publicly available depression dataset proposed by Shen et al. (156). The tweets were crawled and labelled by the authors. The dataset contains three components: **(1) Depressed dataset D1**, which comprises of 2558 samples labelled as depressed users and their tweets, **(2) Non-depressed dataset D2**, which comprises of 5304 labelled non-depressed users and their tweets. **(3) Depression Candidate dataset D3**. The authors constructed a large-scale unlabeled depression-candidate dataset of 58810 samples. In our experiments, we used the labelled dataset: **D1** and **D2**. We preprocess the dataset by excluding users who have their posting history comprising of less than ten posts or users with followers more than 5000, or users who tweeted in other than English so that we have sufficient statistical information associated with every user. We have thus considered 4208 users (51.30% depressed and 48.69 % non-depressed users) as shown in Table 7.1. For evaluation purpose, we split the dataset into training (80%) and test (20%) sets.

3.4.3 Experimental Settings

We have reported our experimental results after performing five-fold cross-validation. For summarization classification model (CNN-BiGRU with attention), the convolution layer, the window of size is set as 3, and the poling size for the max-pooling layer is set as 4. For BiGRU layer, we set the hidden layer to 32. For user behaviours representation model (stacked BiGRU), we used two layers of bidirectional GRU, and we set hidden neurons of each layer to 64. All models are trained using the Adam optimizer (78) using the default parameters: β_1 = 0.9, β_2=0.999, epsilon = 1e-7 and the learning rate = 0.001.

Table 3.3: Effectiveness comparison different methods to detect depression via user behaviours.

Model	Prec.	Rec.	F1	Acc.
SVM	0.724	0.632	0.602	0.644
NB	0.724	0.623	0.588	0.636
MDL	0.790	0.786	0.786	0.787
GRU	0.743	0.705	0.699	0.714
BiGRU	0.787	0.788	0.760	0.750
Stacked BiGRU	**0.825**	**0.818**	**0.819**	**0.821**

We implement the extractive summarization through the clustering model based on Bert Extractive Summarizer [4]. The number of topics in the topic model was set to 5 and from each topic 5 top words were chosen, which gave an overall better performance in our experiments after tuning. During the training phase, we set the batch size to 16. To evaluate the performance of the models, we have used accuracy, precision, recall, and F1 measure. In our framework, the depressed user classification was performed using dual-phased hybrid deep learning model (BiGRU + CNN-BiGRU with Attention)[5] for two attributes user behaviours and user posts summarization. To study the impact of different components in our model, we conduct ablation analysis.

3.4.4 Results

We evaluate the performance of user behaviour (Section 3.3.2) using stacked BiGRU model. We have used four different feature-types, social network, emotions, depression domain-specific and topic modelling, excluding user profile feature and visual feature due to missing values. Table 3.3 shows the comparative results. We notice that behavioural features play an important role in the classification of depression. MLD achieved the second-best performance after Stacked BiGRU; however, stacked BiGRU performs the best to classify diverse features of user behaviours on social media. It outperforms the MLD model in precision by 4%, recall, F1-score and accuracy, by 3%, 3% and 3%, respectively.

For extractive summarization, we have used BERT and performed k-means clustering to select important tweets. We then applied the distilled version of BART (DistilBART) to extract abstract representation from these selected tweets. We used Distilbart provided by Hugging Face[6]. To conduct abstractive summarization, we used different architectures such as XLNet, BERT, RoBERTa, BiGRU with attention and CNN-BiGRU with attention.

[4]https://github.com/dmmiller612/bert-extractive-summarizer

[5]https://github.com/hzogan/DepressionNet

[6]`https://huggingface.co/transformers/model_doc/distilbert.html`

Table 3.4: Comparison of different models for summarization sequence classification.

Model	Prec.	Rec.	F1	Acc.
BiGRU (Att)	0.861	**0.843**	0.835	0.837
CNN (Att)	0.836	0.829	0.824	0.824
CNN-BiGRU (Att)	0.868	0.842	0.833	0.835
XLNet (base)	0.889	0.808	**0.847**	**0.847**
BERT (base)	0.903	0.770	0.831	0.837
RoBERTa (base)	**0.941**	0.731	0.823	0.836

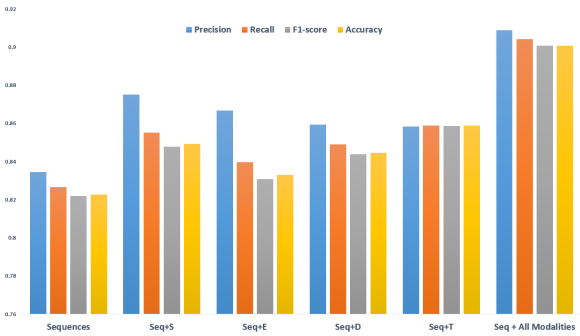
Table 3.5: Comparison of depression detection performances in social media whence of four selected features.

Feature	Model	Precision	Recall	F1-score	Accuracy
User Behaviours	SVM ((132))	0.724	0.632	0.602	0.644
	NB ((132))	0.724	0.623	0.588	0.636
	MDL ((156))	0.790	0.786	0.786	0.787
	GRU ((28))	0.743	0.705	0.699	0.714
	BiGRU	0.787	0.788	0.760	0.750
	Stacked BiGRU	0.825	0.818	0.819	0.821
posts + Image	GRU + VGG-Net + COMMA ((58))	0.900	0.901	0.900	0.900
Posts Summarization	XLNet (base) ((209))	0.889	0.808	0.847	0.847
	BERT (base) ((96))	0.903	0.770	0.831	0.837
	RoBERTa (base) ((96))	**0.941**	0.731	0.823	0.836
	BiGRU (Att)	0.861	0.843	0.835	0.837
	CNN (Att)	0.836	0.829	0.824	0.824
	CNN-BiGRU (Att)	0.868	0.843	0.848	0.835
Summarization + User Behaviures	CNN + BiGRU	0.880	0.866	0.860	0.861
	BiGRU (Att) + BiGRU	0.896	0.885	0.880	0.881
	CNN-BiGRU (Att) + BiGRU	0.900	0.892	0.887	0.887
	BiGRU (Att) + Stacked BiGRU	0.906	0.901	0.898	0.898
	CNN (Att) + Stacked BiGRU	0.874	0.870	0.867	0.867
	DepressionNet (Our Model)	0.909	**0.904**	**0.912**	**0.901**

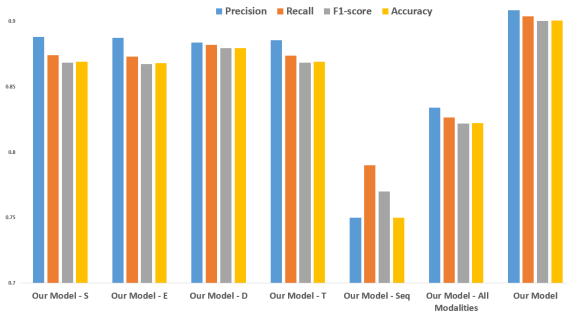
From Table 3.4, we see that BiGRU-Att outperforms other models w.r.t recall, however, XLNet performs best among all the other models w.r.t accuracy and F1-score.

Tables 3.3 and 3.4 show that using only user posts summarization achieves a significantly better performance than using behaviours online. Consequently, using all attributes together further increases the performance of depression detection.

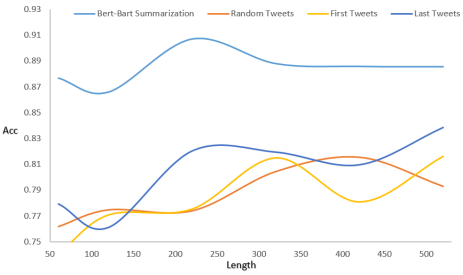
We have combined both summarization and user behaviour representation to capture various features using a hybrid deep learning model. Table 3.5 shows that our DepressionNet (Stacked_BiGRU + CNN-BiGRU_Att) model that show the best results compared with other models. Therefore, our proposed model can effectively leverage online user behaviour and their posts summarization attributes for depression detection. We further conducted a series of experiments and shown the performance of our model considering both qualitative and quantitative analyses. We have also compared the performance



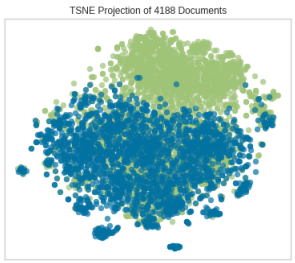
(a) - Effectiveness of summarization sequence with features



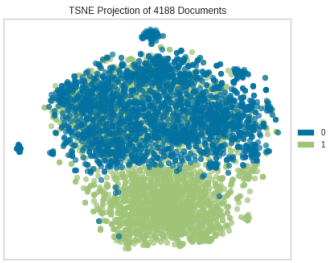
(b) - Effectiveness of our model with features



(c) - Model vs Text length.



(d) -BART



(e) - DistilBART

Figure 3.5: Experimental results: (a) Comparing the performance of our model by concatenating sequences with different feature types. (b) Showing the performance of our model by omitting different features. (c) our model performance vs Text length, with different inputs of data. (d) and (e) T-SNE visualization of Bart and DistilBart Summarization, where 1 represent depressed users posts and 0 represent non-depressed users posts

of our proposed framework with the state-of-the-art methods such as those based on user behaviour as shown in Table 3.3, post summarization as shown in Table 3.4 and user selection of posts with images. Table 3.5 describes the comparative quantitative results. We can observe that the proposed framework DepressionNet performs better in comparison with comparative methods.

DepressionNet outperforms and improves by a percentage of (↑ 6.4%) on depression classification performance and achieves a score of 0.912 (F1-score) in comparison with comparative methods considering both user behaviour and post summarization which is shown in Table 3.5. Gui et al., (58) showed that using images along with user posts plays an important role in depression detection. Similarly, the content of URLs may also be very helpful in the detection of depression as the patient often shares the link about the disease, and its medications among others. In the future, we plan to investigate the images as well as URLs along with post summarization and user behaviour. Our model

also showed a slightly better performance in comparison to the model proposed in (58). The improvement mainly comes from the interplay between the images the depressed user tweeted learnt jointly with text which helps give extra multi-modal knowledge to the computational model. Table 3.5 further shows comparative analysis. We can observe that not only DepressionNet but individual components of the proposed framework show a better performance in comparison to comparative methods. It is noticeable that our stacked BiGRU utilising the behavioural features, the social network, emotions, depression domain-specific and topic modelling outperformed the comparative methods based on leveraging just user behavioural patterns. Stacked BiGRU achieves a score of 0.819 (F1-score) in comparison to MDL and BiGRU. Table 3.3 shows the detailed comparison of methods that comprise of the user behavioural patterns. We can derive similar conclusion considering accuracy, precision and recall. Similarly, our proposed extractive-abstractive summarization with BiGRU-attention showed better performance with 0.848 (F1-score) in comparison to 0.847, 0.831, 0.823, 0.835, 0.824 by XLNet, BERT, RoBERTa, BiGRU with attention and CNN with attention. As depression is a gradually deteriorating disease, we considered the user historical tweets to better understand the user behaviour. As the depressed status intensifies, we can observe the evolving trajectory, however, such minor depression reflection could not be easily detected at early stage of depression, in comparison advanced stage. The selection of depression-related tweets and deep analysis of user history helps diagnose the patient at early stage. Based on experimental results, we have the following **observations**.

1. The consideration of user temporal tweet history helps detect the patient suffering from early-stage of depression.

2. Extractive summarization selecting only important tweets results in not only improving the detection performance but also considerably reducing the computational complexity by filtering out unrelated tweets.

3. The hierarchical cascaded temporal-aware network coalesces multiple layers by integrating the user behavioural representation along with user tweets summary (history-aware temporal post summarization) which further can be used to analyze the progression of the disease and plan managements.

To further analyze the role played by each feature-type and contribution of the user behavioural attributes and user post summarization attribute, we removed the social network feature and denote this model as *Our Model - S*, emotion feature and

denote as *Our Model - E*, domain-specific feature and denote as *Our Model - D* and topic feature which we denote as *Our Model - T*. We have also studied the performance of each attribute from our hybrid model separately. Specifically, we first removed the user behavioural multi-feature attribute (*Our Model - All Feature*) and then removed user post summarization attribute (*Our Model - seq*). We can see in Figure 3.5(b) that our model performance deteriorates as we remove all the multi-features representing user behaviour attributes and degrades more without the post summarization. To understand the effectiveness of sequence attribute, we denote them respectively as following: *Seq+S*, *Seq+E*, *Seq+D* and *Seq+T*. As shown in Figure 3.5(a), the combined sequence with topic feature contributes to depression detection slightly better than other features.

To further analyze the effectiveness of proposed summarization, we have also examined the DepressionNet model with four different data inputs such as first m tweets ($U_i^F = [T_1, ..., T_{20}]$), last m tweets ($U_i^L = [T_{n-20}, ..., T_n]$) and random m tweets ($U_i^R = Random_{20}[T_1, ..., T_n]$). We can notice the significant superiority and stability of summarization approach over different input data as shown in figure 3.5(c). In addition, we have also used t-SNE visualization to evaluate the abstractive summarization. We observe from Figure 3.5(d) and (e) that we can see the distinct separation between the depressed summarization documents and non-depressed for both BART and DistilBART.

3.5 Summary

We have proposed a novel hierarchical deep learning network that coalesces multiple fully connected layers to integrate user behavioural representation and user posting (history-aware posting temporal network). Our framework for depression detection works on online social media data which is characterised by first summarising relevant user post history to automatically select the salient user-generated content. Automatic summarization leads to a natural advantage to our computational framework which enables it to focuses only on the most relevant information during model training which significantly helps reduce the curse-of-dimensionality problem. Automatic summarization also introduces an additional benefit that there are no arbitrary design choices underlying our feature selection, e.g., discarding sentences with with certain predefined words or sentences within a certain length. As a result, our CNN-GRU model learns more discriminative attributes from data than the comparative models. The interplay between automatic summarization, user behavioural representation and model training helps us achieve significantly better results than existing state-of-the-art baselines.

CHAPTER 4

EXPLAINABILITY FOR DEPRESSION DETECTION

4.1 Background and Motivation

Mental illness is a serious issue faced by a large population around the world. In the United States (US) alone, every year, a significant percentage of the adult population is affected by different mental disorders, which include depression mental illness (6.7%), anorexia and bulimia nervosa (1.6%), and bipolar mental illness (2.6%) (8). Sometimes mental illness has been attributed to the mass shooting in the US (105), which has taken numerous innocent lives. One of the common mental health problems is depression that is more dominant than other mental illness conditions worldwide (217). Diagnosis of depression is usually a difficult task because depression detection needs a thorough and detailed psychological testing by experienced psychiatrists at an early stage (145) and it requires interviews, questionnaires, self-reports or testimony from friends and relatives. Moreover, it is very common among people who suffer from depression that they do not visit clinics to ask help from doctors in the early stages of the problem (229).

Individuals and health organizations have shifted away from their traditional interactions, and now meeting online by building online communities for sharing information, seeking and giving the advice to help scale their approach to some extent so that they could cover more affected populations in less time. Besides sharing their mood and actions, recent studies indicate that many people on social media tend to share or give advice on health-related information (60; 116; 153; 137). These sources provide the po-

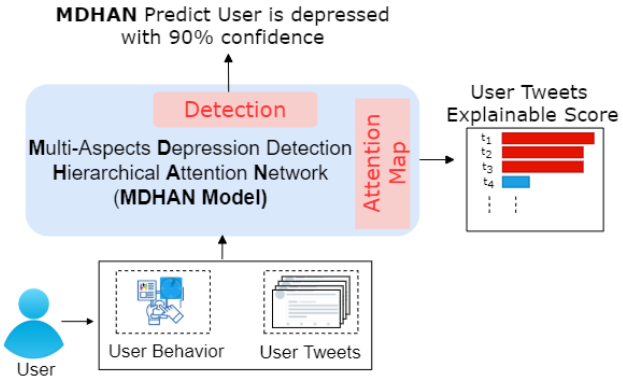


Figure 4.1: Explainable depression detection

tential pathway to discover the mental health knowledge for tasks such as diagnosis, medications and claims. It is common for people who suffer from mental health problems too often "implicitly" (and sometimes even "explicitly") to disclose their feelings and their daily struggles with mental health issues on social media as a way of relief (129; 12). Therefore, social media is an excellent resource to automatically discover people who are depressed. While it would take a considerable amount of time to manually sift through individual social media posts and profiles to locate people going through depression, automatic scalable computational methods could provide timely and mass detection of depressed people which could help prevent many major fatalities in the future and help people who genuinely need it at the right moment. Usually, depressed users act differently when they are on social media, producing rich behavioural data, which is often used to extract various features. However, not all of them are related to depression.

Recently, deep learning has been successfully applied to several application problems, such as stock market predictions (120; 190), traffic flow and traffic accident risk predictions (48; 192; 175), and mental illness detections (76). Moreover deep learning has been applied for depression detection on social media and showed significantly better performance than traditional machine learning methods. Hamad et. al. (224) presented a computational framework for automatic detection of the depressed user that initially selects relevant content through a hybrid extractive and abstractive summarization strategy on the sequence of all user tweets leading to a more fine-grained and relevant content, which then is forwarded to deep learning framework comprising of unified learning machinery of the convolutional neural network coupled with attention-enhanced gated recurrent units leading to better empirical performance than existing strong baseline methods. Even though recent work showed the effectiveness of deep learning methods for depression detection, most of the existing machine learning

methods provide no explainability for depression prediction, hence their predictions are obscure to humans which reduces the trust in the deep learning models. An explainable model provides insights into how a deep learning model can be improved and supports understanding. Thus, to engenders the appropriate user trust and provide the reason behind the decision, we aim to develop an explainable deep learning-based solution for depression detection by utilizing multi-aspect features from the diverse behaviour of the depressed user in social media. Apart from the latent features derived from lexical attributes, we notice that the dynamics of tweets, i.e. tweet timeline provides a crucial hint reflecting depressed user emotion change over time. To this end, we propose a hybrid model, **M**ulti-aspect **D**epression Detection **H**ierarchical **A**ttention **N**etwork **MDHAN** to boost the classification of depressed users using multi-aspect features and word embedding features. Figure 4.1 illustrate the effectiveness of explainability in improving user trust. The model can derive new deterministic feature representations from training data and produce superior results for detecting depression-level of Twitter users, and derive explanations from a user posts content. Besides, we also studied the performance of our model when we used the two components of user posts and his multi-aspect features separately. We found that model performance deteriorated when we used only multi-aspect features. We further show when we combined the two attributes, our model led to better performance. Our model is based on explainable depression detection, which can learn explainable information from a user's tweets. The attention map in Figure 4.1 returns a user's tweets with explainable scores where the higher the score, the more likely tweet that is important and contributed to depression classification. To summarize, our study makes the following **key contributions**:

1. A novel explainable depression detection framework using deep learning of the textual, behavioural, temporal, and semantic aspect features from social media. To the best of our knowledge, this is the first work of using multi-aspect of topical, temporal and semantic features jointly with word embeddings in deep learning for depression detection.

2. Ibuilding a pipeline aided with explainability based on hierarchical attention networks to explain the prediction of depression detection.

3. Extensive experiments are conducted on benchmark depression twitter dataset, which shows the superiority of our proposed method when compared to baseline methods.

4.2 Preliminary on Explainability for Depression Detection

In this section, we will discuss closely related literature and mention how they are different from our proposed method. In general, just like our work, most existing studies focus on user behaviour to detect whether a user suffers from depression or any mental illness. We will also discuss other relevant literature covering word embeddings and hybrid deep learning methods which have been proposed for detecting mental health from online social networks and other resources including public discussion forums.

Understanding depression on online social networks could be carried out using two complementary approaches which are widely discussed in the literature, and they are:

- Post-level behavioural analysis
- User-level behavioural analysis

4.2.1 Post-level behavioural analysis

Methods that use this kind of analysis mainly target the textual features of the user post that is extracted in the form of statistical knowledge such as those based on count-based methods (86). These features describe the linguistic content of the post which are discussed in (36; 67). For instance, in (36) the authors propose a classifier to understand the risk of depression. Concretely, the goal of the paper is to estimate that there is a risk of user depression from their social media posts. To this end, the authors collect data from social media for a year preceding the onset of depression from user profiles and distil behavioural attributes to be measured relating to social engagement, emotion, language and linguistic styles, ego network, and mentions of antidepressant medications. The authors collect their data using crowd-sourcing tasks, which is not a scalable strategy, on Amazon Mechanical Turk. In their study, the crowd workers were asked to undertake a standardized clinical depression survey, followed by various questions on their depression history and demographics. While the authors have conducted thorough quantitative and qualitative studies, they are disadvantageous in that it does not scale to a large set of users and does not consider the notion of text-level semantics such as latent topics and semantic analysis using word embeddings. Our work is both scalable and considers various features which are jointly trained using a novel hybrid deep learning model using a multi-aspect features learning approach. It harnesses high-performance Graphics Processing Units (GPUs) and as a result, has the potential to scale to large

sets of instances. In Hu et al., (67) the authors also consider various linguistic and behavioural features on data obtained from social media. Their underlying model relies on both classification and regression techniques for predicting depression while our method performs classification, but on a large scale using a varied set of crucial features relevant to this task.

To analyze whether the post contains positive or negative words and/or emotions, or the degree of adverbs (178) used cues from the text, for example, *I feel a little depressed* and *I feel so depressed*, where they capture the usage of the word *"depressed"* in the sentences that express two different feelings. The authors also analyzed the posts' interaction (i.e., on Twitter (retweet, liked, commented)). Some researchers studied post-level behaviours to predict mental problems by analysing tweets on Twitter to find out the depression-related language. In (144), the authors have developed a model to uncover meaningful and useful latent structure in a tweet. Similarly, in (156), the authors monitored different symptoms of depression that are mentioned in a user's tweet. In (157), they study users' behaviour on both Twitter and Weibo. To analyze users' posts, they have used linguistic features. They used a Chinese language psychological analysis system called TextMind in sentiment analysis. One of the interesting post-level behavioural studies was done by (156) on Twitter by finding depression relevant words, antidepressants, and depression symptoms. In (139) the authors used post-level behaviour for detecting anorexia; they analyze domain-related vocabulary such as anorexia, eating disorder, food, meals and exercises.

4.2.2 User-level behaviours

There are various features to model users in social media as it reflects overall behaviour over several posts. Different from post-level features extracted from a single post, user-level features extract from several tweets during different times (178). It also extracts the user's social engagement presented on Twitter from many tweets, retweets and/or user interactions with others. Generally, posts' linguistic style could be considered to extract features (67; 213). The authors in (156) extracted six depression-oriented feature groups for a comprehensive description of each user from the collected data set. The authors used the number of tweets and social interactions as social network features. For user profile features, they have used user shared personal information in a social network. Analysing user behaviour looks useful for detecting eating disorders. In Wang et al., (194) they extracted user engagement and activities features on social media. They have extracted linguistic features of the users for psychometric properties which

resembles the settings described in (139; 157) where the authors have extracted 70 features from two different social networks (Twitter and Weibo). They extracted features from a user profile, posting time and user interaction features such as several followers and followee. Similarly, Wong et al. combined user-level and post-level semantics and cast their problem as multiple instances learning setups. The advantage that this method has is that it can learn from user-level labels to identify post-level labels (202).

Recently, Lin et al. (41) applied a CNN-based deep learning model to classify Twitter users based on depression using multi-modal features. The framework proposed by the authors has two parts. In the first part, the authors train their model in an offline mode where they exploit features from Bidirectional Encoder Representations from Transformers (BERT) and visual features from images using a CNN model. The two features are then combined, just as in our model, for joint feature learning. There is then an online depression detection phase that considers user tweets and images jointly where there is a feature fusion at a later stage. In another recently proposed work (25), the authors use visual and textual features to detect depressed users on Instagram posts than Twitter. Their model also uses multi-modalities in data, but keep themselves confined to Instagram only. While the model in (90) showed promising results, it still has a certain disadvantage. For instance, BERT vectors for masked tokens are computationally demanding to obtain even during the fine-tuning stage, unlike our model which does not have to train the word embeddings from scratch. Another limitation of their work is that they obtain sentence representations from BERT, for instance, BERT imposes a 512 token length limit where longer sequences are simply truncated resulting in some information loss, where our model has a much longer sequence length which we can tune easily because our model is computationally cheaper to train. We have proposed a hybrid model that considers a variety of features, unlike these works. While we have not specifically used visual features in our work, using a diverse set of crucial relevant textual features is indeed reasonable than just visual features. Of course, our model has the flexibility to incorporate a variety of other features including visual features.

Multi-modal features from the text, audio, images have also been used in (221), where a new graph attention-based model embedded with multi-modal knowledge for depression detection. While they have used the temporal CNN model, their overall architecture has experimented on small-scale questionnaire data. For instance, their dataset contains 189 sessions of interactions ranging between 7-33min (with an average of 16 min). While they have not experimented with their method with short and noisy data from social media, it remains to be seen how their method scales to such large collections. Xezonaki et al.,

(206) propose an attention-based model for detecting depression from transcribed clinical interviews than from online social networks. Their main conclusion was that individuals diagnosed with depression use affective language to a greater extent than those who are not going through depression. In another recent work (201), the authors discuss depression among users during the COVID-19 pandemic using LSTM and fastText (107) embeddings. In (161), the authors also propose a multi-model RNN-based model for depression prediction but apply their model on online user forum datasets. Trotzek et al., (176) study the problem of early detection of depression from social media using deep learning where they leverage different word embeddings in an ensemble-based learning setup. The authors even train a new word embedding on their dataset to obtain task-specific embeddings. While the authors have used the CNN model to learn high-quality features, their method does not consider temporal dynamics coupled with latent topics, which we show to play a crucial role in overall quantitative performance. Farruque et al., (49) study the problem of creating word embeddings in cases where the data is scarce, for instance, depressive language detection from user tweets. The underlying motivation of their work is to simulate a retrofitting-based word embedding approach (50) where they begin with a pre-trained model and fine-tune the model on domain-specific data.

Opinions and emotions play an important role in detecting depression in social media product feedback, services, and other topics. The analysis of emotions in users' posts has continued to be one of the leading research directions. Prior researches (36; 220; 170) have investigated how emotions and affective states play a role in people's interactions with technology. Recent research in depression identification has shown that excessive self-focused language and negative emotions are key indicators for detecting depressed people (7; 185). De Choudhury et al. (37) collected a Twitter dataset that included postings from people who had been diagnosed with depression. They studied the sentiment, emotion and linguistic of these tweets. They found interesting differences in the usage of words associated with negative emotions for the depressed user's tweets. Additionally, Twitter data analysis reveals that moms' emotional expression, social engagement and linguistic style of moms who experience postpartum depression alter before their baby is even born (36).

Recent studies have started to target depressed users online, extracting features representing user behaviours and classifying these features into different groups, such as the number of posts, posting time distribution, and several followers and followee. Peng et. al. extracted different features and classified them into three groups, user profile, user behaviour and user text and used multi-kernel SVM for classification (133).

The above-mentioned works have some limitations. They mainly focused on studying user behaviour than taking cues from user-generated content such as the text they share which make it extremely difficult to achieve high performance in classification. These models also cannot work well to detect depressed users at the user level, and as a result, they are prone to incorrect prediction. Our novel approach combines user behaviour with user history posts. Besides, our strategy to select salient content using automatic summarization helps our model only focus on the most important information. Although recent deep learning methods showed significant performance for depression detection, most of the existing models do not explain prediction since explainability and effectiveness could sometimes conflict. The explainable model can provide deep insight into how a deep learning model can be improved and supports understanding. Therefore, to provide some details and explain user tweets or reasons to make a decision functioning clear or easy to understand, we aim to develop an explainable deep learning-based approach for depression detection. Our proposed model utilized multi-aspect features from the diverse behaviour of the depressed user and his posts on social media.

4.2.3 Explainable Deep Learning

Deep neural networks help people make better decisions in various industries by producing more accurate and insightful predictions based on vast amounts of data. However, unlike interpretable machine learning methods (43; 191), deep learning models (DNNs) learned representations are typically not interpretable by humans(44). As a result, understanding the representations acquired by neurones in intermediate levels of DNNs is important to the explanation of deep neural networks (DNNs)(93; 92). Meanwhile, concerns about the nature and operation of the deep neural network's black box have grown, driving an increase in curiosity in deconstructing its essential components and understanding its functions. Therefore, explainability has lately received a lot of attention, owing to the requirement to explain the internal mechanics of a deep learning system (210; 23). Many recent studies have focused on improving the transparency of deep neural networks to be adequately understood and be reliable. Attention-based methods can improve model transparency and have shown to be effective in various Natural Language Processing (NLP) tasks, including entity recognition, machine translation systems and text classification (195; 9). Moreover, for document classification (210) and time series forecasting and classification (188), a variety of approaches for designing explainable neural networks employing attention processes have been investigated. In this chapter, we propose using hierarchical attention to improve depression detection by

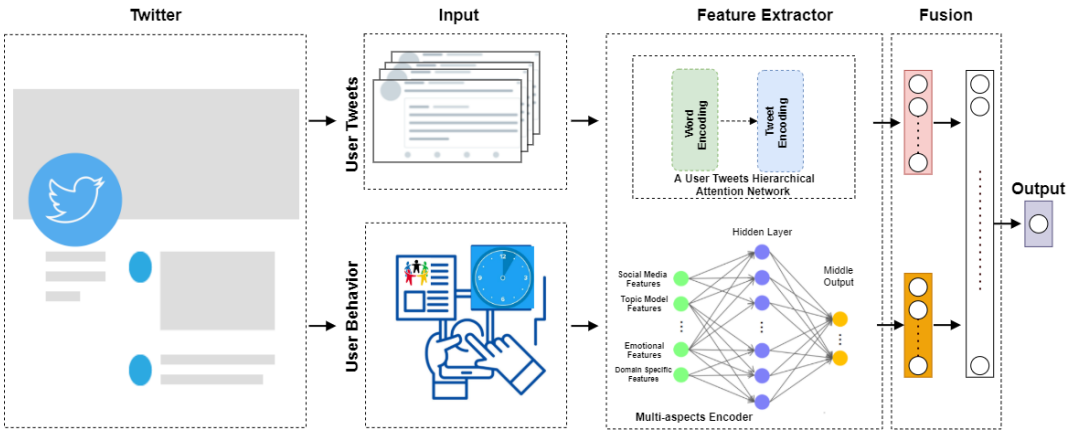


Figure 4.2: Overview of our proposed model MDHAN: We predict depressed user by fusing two kinds of information: (1) User tweets. (2) User Behaviours.

capturing the explainability of depressed user tweets.

4.3 Explainable Deep Depression Detection

Suppose we have a set U of labelled users from both depression or non-depression samples. Let A be a user posts $A = [t_1, t_2, ..., t_L]$ consisting L tweets, where L is the total number of tweets per user, each tweet t_i contains n-words $t_i = [w_{i1}, w_{i2}, ..., w_{iN}]$ where N is the total number of words per tweet. Let M be the features in total for a user $\{m_i\}_{i=1}^{M}$, and let $\{1, 2, ..., S\}$ be a finite set of available aspects features, so we denote M_s as the dimension of S^{th} aspect. Therefore, once we have a user tweets A and a set of related user behaviours feature M. Our depression detection function is represented as follows:

$$f(A, M) \rightarrow \hat{y} \tag{4.1}$$

The model has been designed in such a way that it maximizes prediction accuracy. In our problem, we treat depression detection as the binary classification problem, i.e., user can be depressed ($\hat{y} = 1$) or not-depressed ($\hat{y} = 0$). Due to the complexity of user posts and the diversity of their behaviour on social media, we propose a hybrid model based on Hierarchical Attention Networks (HAN) that combines with Multilayer Perceptron (MLP) to detect depression through social media as depicted in Figure 4.2. For each user, the model takes two inputs for the two attributes. First, the four aspects feature input that represents the user behaviour vector runs into MLP, capturing distinct and latent features and correlation across the features matrix. The second input represents

each user input tweet that will be replaced with its embedding and fed to Hierarchical Attention Networks (HAN) to learn some representation features through a hierarchical word and tweet level encoding. The output in the middle of both attributes is concatenated to represent one single vector feature that fed into an activation layer of sigmoid for prediction. In the following sections, we will discuss the following two existing separate architectures.

4.3.1 Feature Selection

From the depression criteria and online behaviours on social media, we extracted a comprehensive set of depression-oriented features inspired by offline symptoms. Each feature group represents a single aspect. While we did not exploit multimedia features such as images or videos, we used a rich set of features to model multiple aspect. We introduce this attribute type where the goal is to calculate the attribute value corresponding to each features aspect for each user. We mainly consider four major aspects as listed below. These features are extracted respectively for each user as follows:

4.3.1.1 Social Information and Interaction

From this attribute, we extracted several features embedded in each user profile. These are features related to each user account as specified by each feature name. Most of the features are directly available in the user data, such as the number of users following and friends, favourites, etc.

Moreover, the extracted features relate to user behaviour on their profile. For each user, we calculate their total number of tweets, the total length of all tweets and the number of retweets. We further calculate posting time distribution for each user, by counting how many tweets the user published during each of the 24 hours a day. Hence it is a 24-dimensional integer array. To get posting time distribution for each tweet, we extract two digits as hour information, then go through all tweets of each user and track the count of tweets posted in each hour of the day.

4.3.1.2 Emojis Sentiment

Emojis allow users to express their emotions through simple icons and non-verbal elements. It is useful to get the attention of the reader. Emojis could give us a glance at the sentiment of any text or tweet, and it is essential to differentiate between positive and negative sentiment text (122). User tweets contain a large number of emojis which

can be classified into positive, negative and neutral. For each positive, neutral, and negative type, we count their frequency in each tweet. Then we sum up the numbers from each user's tweets to get the sum for each user. So the final output is three values corresponding to positive, neutral and negative emojis by the user. We also consider Voice Activity Detection (VAD) features. These features contain Valance, Arousal and Dominance scores. For that, we count First Person Singular and First Person Plural. Using affective norms for English words, a VAD score for 1030 words are obtained. We create a dictionary with each word as a key and a tuple of its (valance, arousal, dominance) score as value. Next, we parse each tweet and calculate the VAD score for each tweet using this dictionary. Finally, for each user, we add up the VAD scores of tweets by that user, to calculate the VAD score for each user.

4.3.1.3 Topic Distribution

Topic modelling belongs to the class statistical modelling frameworks which help in the discovery of abstract topics in a collection of text documents. It gives us a way of organizing, understanding and summarizing collections of textual information. It helps find hidden topical patterns throughout the process, where the number of topics is specific by the user apriori. It can be defined as a method of finding a group of words (i.e. topics) from a collection of documents that best represent the latent topical information in the collection. In our work, we applied the unsupervised Latent Dirichlet Allocation (LDA) (15) to extract the most latent topic distribution from user tweets. To calculate topic level features, we first consider the corpus of all tweets of all depressed users. Next, we split each tweet into a list of words and assemble all words in decreasing order of their frequency of occurrence, and common English words (stopwords) are removed from the list. Finally, we apply LDA to extract the latent $K = 25$ topics distribution, where K is the number of topics. We have found experimentally $K = 25$ to be a suitable value. While there are tuning strategies and strategies based on Bayesian non-parametric (173), we have opted to use a simple, popular, and computationally efficient approach that helps give us the desired results.

4.3.1.4 Domain-specific features

1- Depression symptom counts: It is the count of depression symptoms occurring in tweets, as specified in nine groups in DSM-IV criteria for a depression diagnosis. The symptoms are listed in Appendix **??**. We count how many times the nine depression symptoms are mentioned by the user in their tweets. The symptoms are specified as a

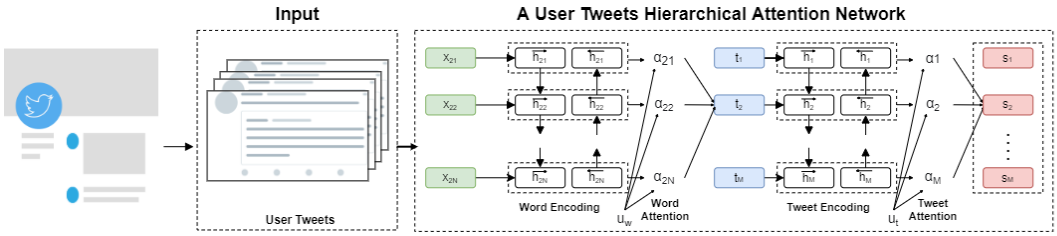


Figure 4.3: An illustration of hierarchical attention network that we used to encode user tweets

list of nine categories, each containing various synonyms for the particular symptom. We created a set of seed keywords for all these nine categories, and with the help of the pre-trained word embedding, we extracted the similarities of these symptoms to extend the list of keywords for each depression symptom. Furthermore, we scan through all tweets, counting how many times a particular symptom is mentioned in each tweet. 2-Antidepressants: We also focused on the antidepressants, and we created a lexicon of antidepressants from the "Antidepressant" Wikipedia page which contains an exhaustive list of items and is updated regularly, in which we counted the number of names listed for antidepressants. The medicine names are listed in Appendix **??**.

4.3.2 User Tweets Encoder Using RNN

Recently, researchers find that the HAN (210; 29) can generate explanations by considering the most important words and sentences in a document. A depressed user could often have different linguistic style posts, including depressive language use, and mentions of antidepressants and symptoms, which can help detect depression. Additionally, a social media post contains linguistic prompts with different levels of word-level and tweet-level. Every word in a tweet and every tweet of a user is equally important to understand a depressed user in social media. For example, "My dad doesn't even seem to believe I'm really hurt!", the word "hurt" contributes more signals to decide whether the tweet is depressed rather than other words in the tweet. So in this way, HAN performs better in predicting the class of given user tweets. Inspired by (210), we proposed Hierarchical Attention Network to learn user tweets representation as depicted in Figure 4.3. We consider U be a user made M tweets $T = [t_1, t_2, ..., T_M]$ each tweet $t_i = [w_1, w_2, ..., w_N]$ contains N_i words. Each tweet is represented by the sequence of d-dimensional embeddings of their words, $t_i = [w_{11}, ..., w_{MN}]$. And we represent each word as the input layer a fixed-size vector from pre-trained word embeddings.

4.3.2.1 Word Encoder

A bidirectional Gated Recurrent Unit (biGRU) is first used as the word level encoder to capture annotations' contextual information. GRU is a Recurrent Nural Network (RNN) that can capture sequential information and sentences' long-term dependency. Only two gate functions are used which are reset and update gates. Update gate has been used to monitor the degree to which the previous moment's status information has been transported into the current state. The higher the update gate value, the more the previous moment's status information is carried forward. The reset gate has been used to monitor the degree to which the previous moment's status information is overlooked. The smaller the reset gate value, the more neglected the context will be. Both the preceding and the following words influence the current word in the sequential textual data, so we use the BiGRU model to extract the contextual features. The BiGRU consists of a forward $\overrightarrow{GRU}$ and a backward $\overleftarrow{GRU}$ that are used, respectively, to process forward and backward data. The annotation w_{ij} represent the word j in a sentence i that contains N-words. Each word of user post (tweet) will convert to a word embedding x_{ij} utilising GloVe (134).

$$\overrightarrow{h_{ij}^{w}} = \overrightarrow{GRU}\left(x_{ij}, \overrightarrow{h_{i(j-1)}}\right), j \in \{1, ..., N\} \tag{4.2}$$

$$\overleftarrow{h_{ij}^{w}} = \overleftarrow{GRU}\left(x_{ij}, \overleftarrow{h_{i(j-1)}}\right), j \in \{N, ..., 1\} \tag{4.3}$$

The combination of the hidden state that is obtained from the forward GRU and the backward GRU $\overrightarrow{h_{ij}^{w}}$ and $\overleftarrow{h_{ij}^{w}}$ is represented as $h_{ij}^{w} = \left[\overrightarrow{h_{ij}^{w}} \oplus \overleftarrow{h_{ij}^{w}}\right]$. Which carries the complete tweet information centred around x_{ij}.

We describe the attention mechanism. It is crucial to introduce a vector u_{ij} for all words, which is trainable and expected to capture global words. The h_{ij}^{w} annotations create the basis for attention that starts with another hidden layer by letting the model learn and randomly initialized biases (b_w) and weights (W_w) through training. the annotations u_{ij} will be represented as follows:

$$u_{ij} = tanh(W_w h_{ij}^{w} + b_w) \tag{4.4}$$

The product $u_{ij}u_w$ (u_w is randomly initialized) expected to signal the importance of the j word and normalized to an importance weight per word α_{ij} by a softmax function:

$$\alpha_{ij} = \frac{\exp(u_{ij}u_w)}{\sum_j \exp(iju_w)} \tag{4.5}$$

Finally, a weighted sum of word representations concatenated with the annotations previously determined called the tweet vector v_i, where α_t indicating importance weight per word:

$$v_i = \sum_t \alpha_{ij} h_{ij}^w \tag{4.6}$$

4.3.2.2 Tweet Encoder

In order to learn the tweet representations h_i^t from a learned tweet vector v_i, we capture the information of context at the tweet level. Similar to the word encoder component, the tweet encoder employs the same BiGRU architecture. Hence the combination of the hidden state that is obtained from the forward GRU and the backward GRU $\overrightarrow{h_i^t}$ and $\overleftarrow{h_i^t}$ is represented as $h_i^t = \left[\overrightarrow{h_i^t} \oplus \overleftarrow{h_i^t}\right]$. Which capture the coherence of a tweet concerning its neighbouring tweets in both directions. Following that, we want to find user tweets that might explain why someone is sad. They should also help identify depression since they give good explainability. Since a user tweets may not be equally important in determining and explaining whether a user is depressed, we use attention over user tweets to capture the semantic affinity of tweets and learn their attention weights based on their relevance to the depression, allowing more reliable and explainable predictions. we will capture the related tweets in the formed vector $\hat{t}$ by using tweet level attention layer. The product $u_i u_s$ is expected to signal the importance of the i tweet and normalized to an importance weight per tweet α_i. Finally, s_i will be a vector that summarizes all the tweet information in a user post:

$$s_i = \sum_t \alpha_i h_i^t \tag{4.7}$$

4.3.3 Multi-Aspect Encoder

Suppose the input which resembles a user behaviour be represented as $[m_1, m_2, ..., m_M\ ]$ where M is the total number of features and M_s is the dimension of S^{th} aspect. Hence, to obtain fine-grained information from user behaviours features, the multi-aspect features are fed through a one-layer MLP to get a hidden representation m_i:

$$p_i = f\left(b + \sum_{i=1}^{M} W_i m_i\right) \tag{4.8}$$

where f stands for the nonlinear function and the outcome of behaviour modelling p_i is the high-level representation that captures the behavioural semantic information and plays a critical role in depression diagnosis.

4.3.4 Classification Layer

At the classification layer, we need to predict whether the user is depressed or not depressed. So far, we have introduced, how we encode user multi aspect behaviours features (p) and how we can encode user tweets by modelling the hierarchical structure from word level and tweet level (s). Then from both components, we construct the feature matrix of user behaviours features and user tweets:

$$p = p_1, p_2, ..., p_M \in \mathbb{R}^{1d \times M} \tag{4.9}$$

$$s = s_1, s_2, ..., s_n \in \mathbb{R}^{2d \times n} \tag{4.10}$$

We further unify these components together, which is denoted as $[p, s]$. The output of such a network is typically fed to a sigmoid layer for classification:

$$\hat{y} = sigmoid(b_f + [p, s]W_f) \tag{4.11}$$

where where $\hat{y}$ is the predicted probability vector with $\hat{y_0}$ and $\hat{y_1}$ indicate the predicted probability of label being 0 (not depressed) and 1 (depressed user) respectively. Then, we aim to minimize the cross-entropy error for each user with ground-truth label y:

$$\text{Loss} = -\sum_i y_i \cdot \log \hat{y}_i$$

where $\hat{y}_i$ is the predicted probability and y_i is the ground truth label (either depression or non-depression) user.

4.3.5 Explainability

We aim to select user tweets that can explain why a user is depressed. As they provide a reasonable explanation, they should also help detect depression. The hierarchical attention that we explained in the previous sections in the word and tweet encoding is a suitable mechanism for giving high weights of user tweets representations. Besides, the explainability degree of user tweets are learned through the attention weight. Since varied words have different weights in each tweet based on the attention map, it indicates that our model can extract important and long-range contextual information from a tweet. Generally, the attention map of our model can select the most contributed words that identify a depressed and their corresponding tweets. Therefore, user tweets with high attention weight are essential and likely explain why a user is depressed.

4.4 Experiments and Results

In this section, we present the experimental evaluation to validate the performance of **MDHAN**. First will we will introduce datasets and evaluation Metrics and experimental settings, followed by the experimental results.

4.4.1 Comparative Methods

We compare our model with the following classification methods:

- **MDL: Multimodal Dictionary Learning Model** is to detect depressed users on Twitter (156). They use dictionary learning to extract latent data features and sparse representations of a user.

- **SVM: Support Vector Machines** is a popular and strong classifier that has been applied on a wide range of classification tasks (75) and it remains a strong baseline.

- **NB: Naive Bayes** is a family of probabilistic algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between instances (117).

- **BiGRU**:We applied **Bidirectional Gated Recurrent Unit** (26) with attention mechanism to obtain user tweets representations, which we then used for user tweets classification.

Description	Depressed	Non-Depressed
Numer of users	2159	2049
Numer of tweets	447856	1349447

Table 4.1: Summary of labelled data used to train MDHAN model

- **MBiGRU**: Hybrid model based on MLP and BiGRU for multi-aspect features for the user behaviour and the user,Äôs online timeline (posts).
- **CNN**: We utilized **Convolutional Neural Networks** (77) with an attention mechanism to model user tweets, which can capture the semantics of different convolutional window sizes for depression detection.
- **MCNN**: Hybrid model based on MLP and CNN for multi-aspect features for the user behaviour and the user,Äôs online timeline (posts).
- **HAN**: A hierarchical attention neural network framework (210), it used on user posts for depression detection. The network encodes first user posts with word-level attention on each tweet and tweet-level attention on each user post.
- **MDHAN**: The proposed model in this chapter.

4.4.2 Datasets

Recent research conducted by Shen et al. (156) is one such work that has collected large-scale data with reliable ground truth data, which we aim to reuse. To exemplify the dataset further, the authors collected three complementary data sets, which are:

- Depression data set: Each user is labelled as depressed, based on their tweet content between 2009 and 2016.
- Non-depression data set: Each user is labelled as non-depressed and the tweets were collected in December 2016.
- Depression-candidate data set: The authors collected are labelled as depression-candidate, where the tweet was collected if contained the word “depress”.

Data collection mechanisms are often loosely controlled, impossible data combinations, for instance, users labelled as depressed but have provided no posts, missing values, among others. After data has been crawled, it is still not ready to be used directly by the machine learning model due to various noise still present in data, which is called the

"raw data". The problem is even more exacerbated when data has been downloaded from online social media such as Twitter because tweets may contain spelling and grammar mistakes, smileys, and other undesirable characters. Therefore, a pre-processing strategy is needed to ensure satisfactory data quality for computational modal to achieve reliable predictive analysis.

To further clean the data we used Natural Language processing ToolKit (NLTK). This package has been widely used for text pre-processing (64) and various other works. It has also been widely used for removing common words such as stop words from text (39; 144). We have removed the common words from users tweets (such as "the", "an", etc.) as these are not discriminative or useful enough for our model. These common words sometimes also increase the dimensionality of the problem which could sometimes lead to the "curse-of-dimensionality" problem and may have an impact on the overall model efficiency. To further improve the text quality, we have also removed non-ASCII characters which have also been widely used in literature (213).

Pre-processing and removal of noisy content from the data helped get rid of plenty of noisy content from the dataset. We then obtained high-quality reliable data which we could use in this study. Besides, this distillation helped reduce the computational complexity of the model because we are only dealing with informative data which eventually would be used in modelling. We present the statistics of this distilled data below:

- Number of users labelled positive tweets: 5899.
- Number of tweets from positive users: 508786.
- Number of users labelled negative: 5160.
- Number of tweets from negative users: 2299106.

To further mitigate the issue of sparsity in data, we excluded those users who have posted less than ten posts and users who have less than 5000 followers, therefore we ended up with 2159 positive users and 2049 negative users.

For our experiments, we have used the datasets as mentioned in section (3). They provide a large scale of data, especially for labelled negative and candidate positive, and in our experiments, we used the labelled dataset. We preprocess the dataset by excluding users who have their posting history comprising of less than ten posts or users with followers more than 5000, or users who tweeted in other than English so that we have

Matric	SVM	NB	MDL	BiGRU	MBiGRU	CNN	MCNN	HAN	**MDHAN**
Accuracy	0.644	0.636	0.787	0.764	0.786	0.806	0.871	0.844	**0.895**
Precision	0.724	0.724	0.790	0.766	0.789	0.817	0.874	0.870	**0.902**
Recall	0.632	0.623	0.786	0.762	0.787	0.804	0.870	0.840	**0.892**
F1-score	0.602	0.588	0.786	0.763	0.786	0.803	0.870	0.839	**0.893**

Table 4.2: Performance comparison of MDHAN against the baselines for depression detection on (156) Dataset

sufficient statistical information associated with every user. We have thus considered 4208 users (51.30% depressed and 48.69 % non-depressed users) as shown in Table 7.1. For evaluation purposes, we split the dataset randomly into training (80%) and test (20%), and we have reported our experimental results after performing five fold cross-validation.

4.4.3 Experimental Setting and Evaluation Metrics

For parameter configurations, the word embeddings are initialized with the Glove (134) with a dimension of 100 on the training set of each dataset to initialize the word embeddings of all the models, including baselines. The hidden dimension has been set to 100 in our model and other neural models, also, the dropout is set to 0.5. All the models are trained to use use the Adam optimization algorithm (79) with a batch size of 16 and an initial learning rate of 0.001. Then we trained our model for 10 iterations, with a batch size of 16. The number of iterations was sufficient to converge the model and our experimental results further cement this claim where we outperform existing strong baseline methods, and the training epoch is set to 10. We used python 3.6.3 and Tensorflow 2.1.0 to develop our implementation. We rendered the embedding layer to be not trainable so that we keep the features representations, e.g., word vectors and topic vectors in their original form. We used one hidden layer and a max-pooling layer of size 4 which gave a better performance in our setting. Finally, we employ traditional popular metrics such as precision, recall, F1, and accuracy.

4.4.4 Experimental Results

In our experiments, we study our model attributes including the quantitative performance of our hybrid model. For the multi-aspect features and user's timeline semantic features attribute, we will use both these attributes jointly. After grouped user behaviour in social media into a multi-aspect attribute, we evaluate the performance of the model.

First, we examine the effectiveness of using the multi-aspect features only for depression detection with different classifiers. Second, we showed how the model performance increased when we utilize multi-aspect features with hierarchical attention network MDHAN. We summarise the results in Table 4.2 as follows:

- Naive Bayes obtain the lowest F1 score, which demonstrates that this model has less capability to classify tweets when compared with other existing models to detect depression. The reason for its poor performance could be that the model is not robust enough to sparse and noisy data.
- MDL model outperforms SVM, NB and BiGRU, and obtains better accuracy than these three methods. Since this is a recent model specially designed to discover depressed users, it has captured the intricacies of the dataset well and learned its parameters faithfully leading to better results.
- we can observe the evolving when we integrate The multi-aspect features with user posts and that better helped to analyze a user that seems to be depressed as shown in the performance of MBiGRU, MCNN MDHAN.
- We can see our proposed model MDHAN improved the depression detection up to 10% on F1-Score, compared to MDL model and 5% compared to HAN model. This suggests that our model outperforms a strong model. The reason why our model performs well is primarily that it leverages a rich set of features which is jointly learned in the estimation of the consolidated parameters resulting in a robust model.
- Furthermore, MDHAN achieved the best performance with 89% in F1, indicating that combining HAN with multi-aspect strategy for user timeline semantic features strategy is sufficient to detect depression in Twitter. We can also deduce from the table that our model consistently outperforms all existing and strong baselines.

4.4.5 Comparison and Discussion

To get a better look at our model performance, We have compared the effectiveness of each of the two attributes of our model. Therefore, we test the performance of the model with a different attribute, we build the model to feed it with each attribute separately and compare how the model performs. First, we test the model using only the multi-aspect attribute, we can observe in Fig **??** the model perform less optimally when we used

MLP for **M**ulti-aspect features (MM). In contrast, the model performs better when we use only HAN with word embedding attributes. This signifies that extracting semantic information features from user tweets is crucial for depression detection. Thus, we can see the MDHAN model performance increased when combined both MM and HAN, and outperforms when using each attribute independently. One of the key parameters in MDHAN is the number of tweets for each user; we eventually observed that MDHAN reached optimal performance when using 200 tweets as the maximum number of tweets. Figure 5.9 illustrates the performance of our model concerning the number of tweets.

To further analyze the role played by each aspect features and contribution of the user behavioural attributes and user posts attribute, we removed the four aspects separately as following: the domain-specific feature and denote as *MDHAN - D*, emotion feature and denote as *MDHAN - E*, the social network feature and denote this model as *MDHAN - S* and topic feature which we denote as *MDHAN - T*. We can see in Figure 5.10 that our model performance deteriorates as we remove the topic feature from the MDHAN model and degrades more without the social network features. To dive deeper and understand the effectiveness of each aspect, we combine each aspect separately with HAN and denote them respectively as following: *D+HAN*, *E+HAN*, *S+HAN* and *T+HAN*. As shown in Figure 4.7, we could see that MDHAN with four aspects outperforms the others, which means that each aspect does contribute to depression detection.

4.4.6 Case Study

To illustrate the importance of MDHAN for explaining depression detection results, we visualize the attention map for an example of a depressed user to show the words and tweets captured by MDAHN in Figure 4.8. The words and tweet weights are indicated by the red in this example, the words and tweet are more important by attention weight if the colour is darker. Varied words have different weights in each tweet based on the attention map. It indicates that our model can extract important and long-range contextual information from a tweet. Generally, the attention map of our model can select the most contributed words that identify a depressed user, like mental, patients, therapies and illness, and their corresponding tweets. Tweets containing some words that have not contributed to classifying a depressed user and low attention weight will be neglected, for example, in the figure, we will notice that the first tweet has got the most attention, and the same goes for the words: mental and illness that had the highest weights when determining the prediction of class depression. The figure demonstrates that the attention map gives higher weights to explainable depression

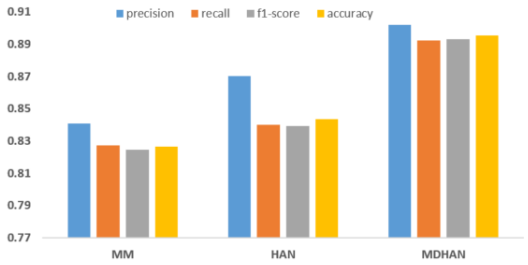


Figure 4.4: Effectiveness comparison between MDHAN with different attributes.

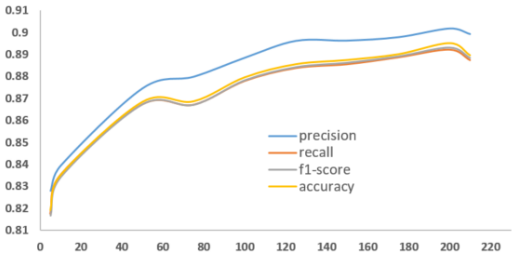


Figure 4.5: Model vs number of tweets

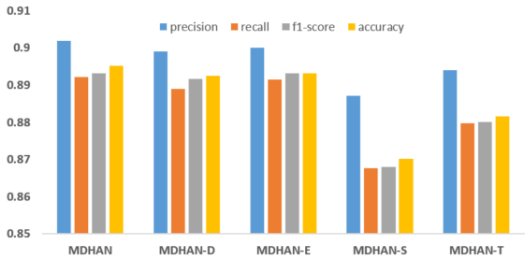


Figure 4.6: Comparisons of various attributes

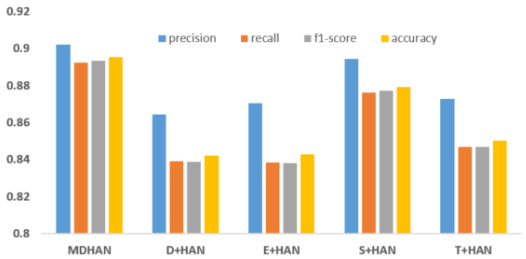


Figure 4.7: Comparison of various use of attributes

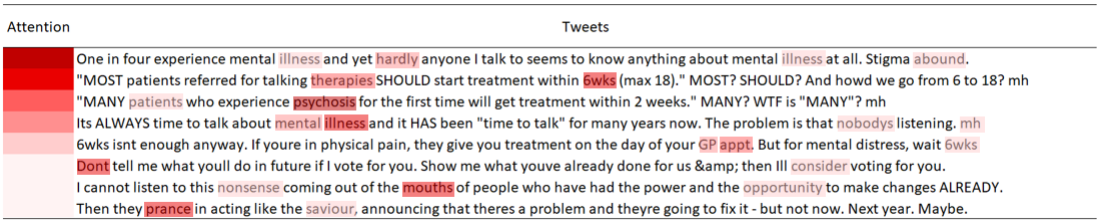


Figure 4.8: Explainability via visualization of attention score in MDHAN

tweets; for instance, the tweet *"One in four experience mental illness ... "* gained the highest attention score among all the user tweets. Moreover, MDAHN can give higher weights to explainable tweets than those interfering and unrelated tweets, which can help select more related tweets and to be a more important feature to detect the depressed user.

To further investigate the five most influencing symptoms among depressed users, we collected all the tweets associated with these symptoms. Then we created a tag cloud (186) for each of these five symptoms, to determine what are the frequent words and importance that related to each symptom as shown in Figure 4.9 where larger font words are relatively more important than rest in the same cloud representation. This cloud gives us an overview of all the words that occur most frequently within each of these five symptoms.

(a) A word cloud for Suicidal ideation.

(b) A word cloud for Worthlessness.

(c) A word cloud for Loss of Energy.

(d) A word cloud for Insomnia.

(e) A word cloud for depression mood.

Figure 4.9: A word cloud depicting the most influencing symptoms.

4.5 Summary

We have proposed explainable Multi-Aspect Depression Detection with Hierarchical Attention Network (MDHAN) for detecting depressed users through social media analysis by extracting features from the user behaviour and the user's online timeline (posts). We have used a real-world data set for depressed and non-depressed users. Our main contribution is a novel hybrid computational model that can not only effectively model the real-world data but can also help derive explanations from them. We assign the multi-aspect attribute which represents the user behaviour into the MLP and user timeline posts into HAN to calculate each tweet and words' importance and capture semantic sequence features from the user timelines (posts). Our model shows that training this hybrid network improves classification performance and identifies depressed users outperforming other strong methods and ensures adequate evidence to explain the prediction. In the future, We will analyze users' tweets by considering topics and sentiments simultaneously to provide supporting evidence for each Depression DSM-IV criteria. Moreover, we will go beyond social media content and use URLs, images, and a

mix of short and long user-generated content with traditional web pages. This would help give more contextual knowledge to the model that will help us focus on a task where our model not only detects depression but also automatically suggests the possible diagnosis.

CHAPTER 5

NarrationDep: Modeling Narrative Elements in Social Media to Identify Depression

5.1 Background and Motivation

Detecting depression via understanding a user's language through social media is challenging. One way that has been used to analyze depressive language use on more general social media channels like Twitter is to utilize statements of diagnosis such as "(I'm/I was/ I am/ I've been) diagnosed depression" (34; 156) to determine what includes a case of mental illness. Furthermore, this approach is limited because a person who openly publishes a diagnosis on Twitter is not someone they are attempting to hide and may not be part of the group of undiagnosed people that such approaches may wish to identify. Furthermore, it may be more possible to tweet about disorder-related issues, leading us to believe that mental health status can be recognized more easily from social media data. For example, Zogan et al, (225), presented a deep network that first summarises the relevant user post history in order to identify the most crucial user-generated tweet automatically for a depressed user; the framework leads to a natural advantage to enable focus only on the most relevant information during model training. However, previous studies still face challenges, and some of them lack a resounding and sufficient analysis of people with depression due to the nature of social media data that is not sequential, unconnected, and unordered. We argue that there is a critical need to understand the

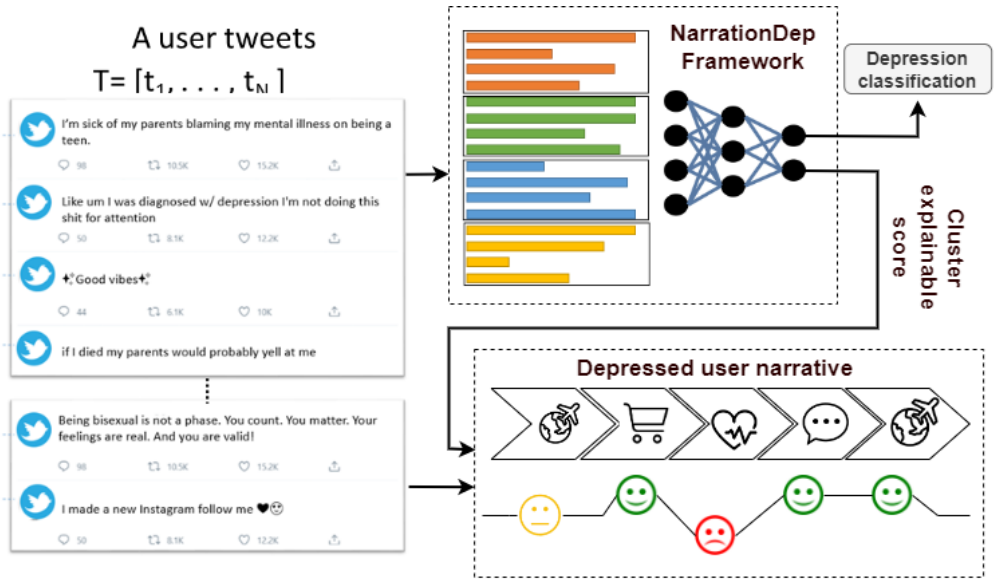


Figure 5.1: NarrationDep is an algorithm module that uses explainable deep learning for depression detection. It provides detection results and explanations about the narrative of a user who is depressed.

content of the depressed person and extract a chain of events to tell a story from a particular point, the topic of these events, and their sentiments when they deal with these events. In other words, we need different information about each user's posts, such as user Narrative posts, in order to extract silent information about each user and to differentiate depressed users from a non-depressed ones.

The narrative plays a very important role in the understanding of events that occur in news or stories (21) or in our daily lives and contains many social as well as moral norms. Narrative in social media refers to the way in which individuals, organizations, and other entities use social media platforms to tell stories and convey information (101; 127). This can include text, images, videos, and other forms of content and can be used for a wide range of purposes, such as building brand awareness, promoting products or services, and engaging with audiences. One important aspect of narrative in social media is the role it plays in shaping people's perceptions and understanding of events and issues (63). Social media narratives can be used to inform, entertain, persuade, and mobilize people around particular issues or causes. They can also be used to shape public opinion, influence political discourse, and even shape public policy. The ability to generate a narrative can enhance the user's experience and make it more engaging and

interactive, and how valuable is the narrative for government, such as defense? User narrative understanding on social media is one of the most captivating approaches to understanding the users' intentions that might be of interest to several organizations such as defense agencies, advertisement agencies and healthcare.

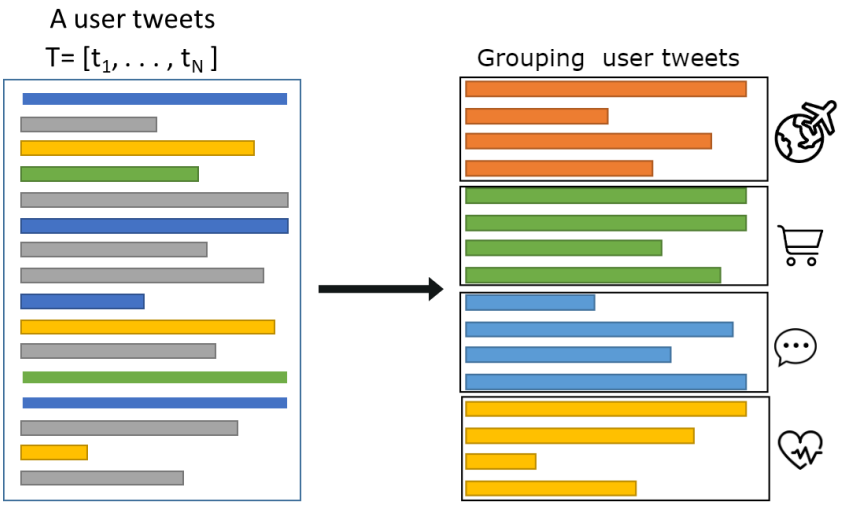


Figure 5.2: Grouping user tweets could provides better understanding and make inferences about personality, relationship, intents, actions, etc

We hypothesize that user posts can tell us a story, but the events of that story are neither sequential nor coherent, which makes it challenging to form it as an ordinary story. Thus, to analyze a user posts story, we need to grasp a user narrative and model narrative elements of user posts. But at first, we must expound on what we imply about the narrative and what it means in social media. Therefore, rather than being the story itself, the narrative is a representation or a particular form of a story. The narrative transforms a story into knowledge or, better yet, into information, and each event is a unit of knowledge that could give us a hint about the story. Therefore events propel the narrative, and the narrative consists of a chain of events.

On social media, people usually engage in events and pursuits that they find meaningful and interesting. Sometimes, the posts about a particular event could be a chronology feed. A chronological feed essentially means a user uses posts in social media to display content via a timeline in a sequential time format. However, many times some events would be shared in different chronologies. So, the social media narrative may be defined as a lengthy story that is divided up into postings for social media platforms like Twitter and Facebook. Even though the posts are brief, they might be combined to create a broader story with a theme Figure 5.2.

Therefore our question is to investigate the feasibility of using social media posts to extract features that can provide a narrative explanation of a series of events and how these features may differ when comparing individuals with a mental disorder. However, the challenging task here, when we have a huge pile of user tweets, and we have to find mysterious patterns you believe are hidden within. So we need a method to organize the collection. It turns out that we can do so by cluster analysis, to aim at partitioning data into coherent groups and find patterns and groupings in our data. So the final output contains a set of clusters each including a set of tweets. Our goal is to understand to what extent social media posts can extract such features to understand and generate a narrative of a series of events. While AI and NLP systems have yet to achieve the goal of learning and generating narratives for a regular text, it will undoubtedly be a major challenge to achieve this aim for social media data such as depressed users' posts on Twitter. Because of the variety of tweets, some of them are not even relevant. If there is any relevance or pattern for these tweets, it requires an understanding of crucial narrative elements and how they evolve and extract a story. Additionally, narrative tweets can be used to identify patterns and trends in the way individuals talk about their mental health as we have shown in (Section 5.4.4), which can help in understanding the experience of depression. It can also be used as a tool for self-expression and self-reflection, which can be beneficial for individuals struggling with depression. Narrative tweets can be a valuable tool for understanding depression, but it's important to remember that tweets are not a substitute for professional diagnosis or treatment. However, extracting narrative tweets can be a way to identify individuals at risk of depression and reach out to them with help.

In this chapter, we address these challenges by creating new tasks aimed specifically towards modeling narrative elements in Social Media. This chapter aims to prove this hypothesis: these groupings will index user tweets based on some imposed scale of similarity and differentiated based on some imposed boundary, which will be beneficial in gaining knowledge about user posts' narrative. We emphasize that what we care about is tweets clustering will be advantageous to identifying depression users and explaining the cause of depression.

Our solutions to these challenges lead in a new framework known as **NarrationDep** (**Narra**tive Detec**tion** for **Dep**ression). The model can generate new deterministic feature representations from training data, perform better at identifying Twitter users who are depressed and generate narrative tweet explanations from user post content. Our model ***NarationDep*** consists of two attributes; one attribute represents user content (tweets

history) by using two attentions (word and tweets level). For the second attribute, we proposed a novel hierarchical attention network for cluster tweets, which we called the Hierarchical Attention Based Clustering Network (HACN). Word, tweet, and cluster tweet levels can be used to generate significant feature representations by our model HACN. We also examined how well our model performed when we individually used the user tweets and content clusters. Furthermore, we demonstrate that our model performed better when we integrated the two attributes. Our model is designed for narrative-based depression detection, which is capable of learning explainable information and extract narrative tweets from a user's content. The user's cluster tweets from the attention map in Figure 5.1 are returned with attention scores, and the higher the score, the more probably it is that the cluster was significant and contributed to the classification of depression, and each tweet in that cluster represent the narrative tweet in user content. Hence, in this chapter, we raise the following research question: Can highlighting meaningful clusters identified by our model ***NarationDep*** help to extract the depressed user narrative elements?

To sum up, our research makes the following **key contributions**:

1. We study a novel problem of modeling narrative elements in social media to analyze how our posts can be used to understand our narrative.

2. We model the social media posts via clustering, short summarization and their intentions as a narrative element in an interactive generation framework. We developed a model-based hierarchical framework for interactively incorporating a user's social media posts to generate a narrative automatically.

3. We provide a principled way to jointly exploit all user contents and cluster tweets to capture an explainable cluster for user narrative understanding with the perspective of depression detection. We have used explainable clusters aided with sentiment analysis to generate the narrative of social media users.

4. Extensive experiments conducted on the benchmark dataset showed significant improvement in depression detection compared to the state-of-the-art methods.

5.2 Preliminary on Narrative in Social Media to Identify Depression

In this section, we summarise some closely related work. We also mention how the model developed in this work is substantially novel and different from our previous works.

5.2.1 Text mining and Narrative

In the field of mental health, text mining has been used to study depression (193; 205; 131). Studies have used text mining to analyze social media posts (113), electronic medical records (205; 53), and other sources of text data to identify patterns and trends in the way people talk about depression (100; 14). Some studies have used text mining to identify markers of depression, such as specific words or phrases (47), while others have used text mining to understand how people talk about depression in different contexts (81) or across different demographic groups (82; 223). Overall, text mining has been used to gain new insights into the way that people experience and talk about depression, which can help mental health professionals better understand the condition and develop more effective treatment strategies. Moreover, text mining and narrative are closely related in the sense that text mining techniques are often used to analyze and understand narrative texts. Narrative text mining specifically applies text mining techniques to extract structured information from narrative texts. This can include identifying key elements of the narrative such as characters, events, and themes, as well as understanding the relationships between these elements. By using text mining techniques to analyze narrative texts, researchers and analysts can gain insights into the text and its meaning that would be difficult to obtain through manual analysis. Additionally, text mining can also be used to extract narrative summaries (166; 54) and narrative classification (56; 219).

In a recent study, Catipon et al. (20) studied the distinctions between conservative narratives on Twitter and found that media bias levels were more pronounced on Parler than on Twitter. They also noticed that subjects that are more controlled on Twitter had vastly different perspectives on Parler, and that well-known news sources were more politically diverse on Parler. Additionally, Hussain et al. (70) created a narrative visualization tool to assist analysts in identifying diverse themes and related narratives being discussed on various blogs. In (148) the authors introduced THEaiTRE 1.0, a system based on GPT-2 for generating theatre play scripts. However, extracting a narrative from

social media, such as Twitter, can be challenging because the text is often unstructured, with a limited number of characters and a high degree of noise and variation in language, making it difficult to extract meaningful information. Narrative tweets, or tweets that tell a story or describe an event or experience, can potentially be a valuable tool for understanding depression. Therefore, in this study, we will present, for the first time, extracting narrative tweets and using them to understand depression. our model is also significantly different from those works mentioned above.

5.2.2 Depression Detection on Social Media

Social media platforms have the potential to assist in identifying and developing methods for diagnosing major depressive disorders. Researchers have been examining the impact of social media on depression prediction as these platforms offer the ability to evaluate an individual's state of mind, thoughts, and content. Studies have demonstrated that analyzing user content and textual information on social media can be effective in detecting depression and other mental illnesses (32; 36; 38).

Wang et al.(194) proposed a method for automatically gathering people who described themselves as having an eating disorder in their Twitter profile descriptions to detect eating disorders within social media communities. The authors collected features of linguistics from users for psychometric qualities, and they used similar settings described in (139; 85; 157). From Twitter and Weibo, the authors collected 70 features. They took these characteristics from a user's profile and user engagement characteristics like many followees and followers. Zogan et al., (227) recently introduced a new model for detecting depressed users using social media, which is an interplay between multilayer perceptron (MLP) and hierarchical attention network (HAN). MLP was used to encode users' online behaviour, while HAN encoded all user tweets at two levels: word-level and tweet-level. They determined each tweet and word weight and extracted characteristics derived from user tweets' semantic sequences. In another work by (224), the authors argued that using all user tweets to identify a depressed user is ineffective and could even degrade a model performance; therefore, they proposed a new summarization framework interplay between extractive and abstractive summarization to generate a shorter representation of user historical tweets and help to reduce the influence of content that may not eventually benefit the classifier.

In contrast to the above, our novel method combines grouping tweets with a user's previous posts to enhance depression identification. Additionally, our strategy of selecting relevant content by using a clustering method enables our model to concentrate solely

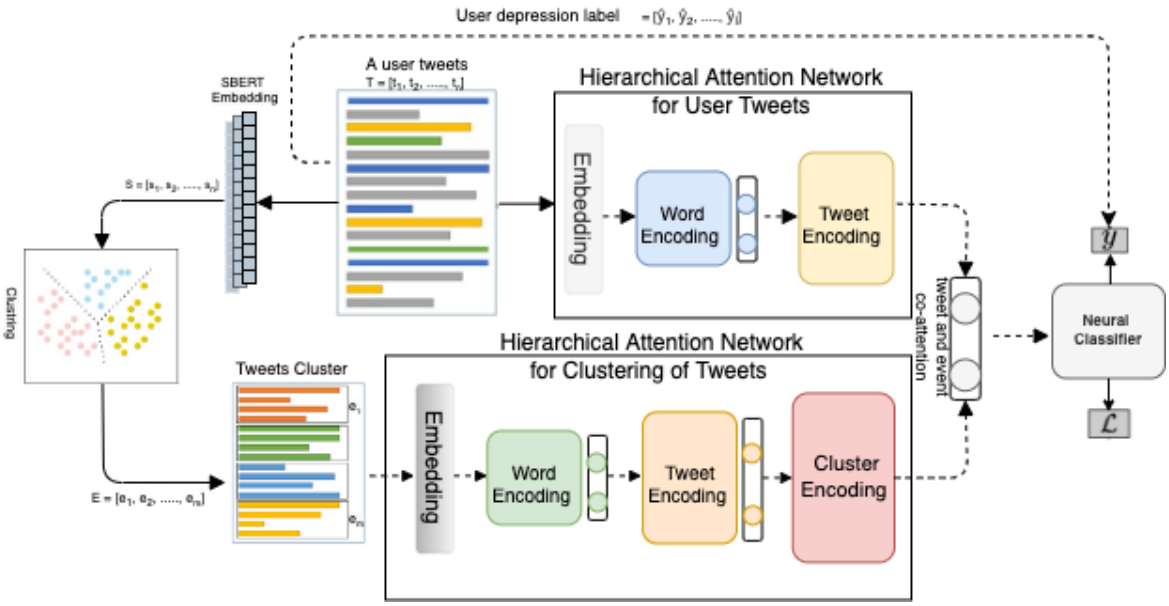


Figure 5.3: An illustration of NarrationDep model

on the most vital information and grasp the narrative element of a depressed user of depressed user.

5.3 Our Novel NarrationDep Framework

5.3.1 High-level Description

We have depicted, in Figure 5.3, a high-level architecture of our novel ***NarationDep*** consisting of different units. We consider user tweets that are in the form of short sentences. We then obtain the vector representation of each of these tweets using the popular SBERT (143) model. We then use a clustering model to semantically cluster tweet representations and obtain the clustered tweets. There are two key units that play a central role to model individual user tweets and those that are clustered together. The reason why we have this design is that we could not only look at individual tweets but we can also model semantically related tweets in a unified network. Then there is a co-attention framework that models the importance of the instances that are fed as features to the neural classifier. Note that the reason why this model has the potential to perform reliably compared to the existing methods is that different components work in a unified way in a single model mitigating error propagation as what is commonly seen in cascaded methods. The co-attention layer not only aids in reliable prediction but also

plays a key role in narrative explanations.

In the user tweets unit, our input is individual user tweets. Given that a user's tweets might be similar due to the homogeneity of user interests, we adopt a clustering approach to cluster tweets into different meaningful clusters. Clustering tweets will reduce the size of event spaces by grouping similar tweets' linguistic information even helping us mitigate the issue of redundancy (224). We cluster tweets by determining a measure of similarity between tweets and combining tweets that are similar into single events, resulting in the elimination of the distinction between the individual tweets that make up the event. We argue that jointly learning a user tweets representations from two different hierarchical structures, a user's contents (words from tweets, tweets from a user's content) and a user's cluster tweets (words from tweets, tweets from a cluster and a cluster from the user's content) at tweet-level and cluster-level encoders are helpful in enhancing depression detection performance. Our framework can jointly learn and extract linguistic relationships by exploiting the local features from user cluster tweets. It can then learn to reliably output explainability and user narratives. We now describe different units of our model in detail.

5.3.2 Hierarchical Attention Network for Clustering of Tweets

Individual user tweets clustering: We develop a framework which learns a low-dimensional representation from potentially ambiguous tweets by users. These clusters can then be applied to predict local themes at the cluster level. By clustering tweets, we generate a more coherent and organized representation of these tweets. As a result, we can see these representations being applied in various downstream applications such as tweets classification. In our framework, during the clustering step, we exploit a technique that groups tweets together and resulting in semantically similar tweets coming close to each other in the semantic space. In particular, we have used UMAP for reducing the dimensionality that outputs low-dimensional representations of tweets and HDSBCAN for clustering. While there are other models that we could use to accomplish our task, we found that the two models we have chosen work reliably on our dataset including meeting our efficiency needs, given our computational resources. The optimal values for the method's parameters are determined through a Bayesian search that is popularly used and is based on a user-defined objective function and constraints.

Word encoding: Our approach employs a bidirectional Gated Recurrent Unit (BiGRU) as the word-level encoder to extract context-sensitive information from the annotations. GRU is a type of Recurrent Neural Network (RNN) that is able to capture sequential

information and the long-term dependencies of sentences. The model only uses two gate functions, namely the reset and update gates. The update gate controls how much of the previous moment's information is carried forward into the current state. A higher update gate value means more of the previous moment's information is retained. The reset gate controls how much of the previous moment's information is ignored. A smaller reset gate value means more of the context is disregarded. Both preceding and following words affect the current word in sequential textual data, so we use the BiGRU model to extract contextual features that take into account both directions. The BiGRU consists of a forward $\overrightarrow{GRU}$ and a backward $\overleftarrow{GRU}$ that are used, respectively, to process forward and backward data. The annotation w_{ijm} represents the word m in a tweet j and cluster i that contains M-words. Each word of a user post (tweet) will convert to a word embedding x_{ijm} utilizing GloVe (134).

$$\overrightarrow{h^w_{ijm}} = \overrightarrow{GRU}\left(x_{ijm}, \overrightarrow{h_{ij(m-1)}}\right), m \in \{1, ..., M\} \tag{5.1}$$

$$\overleftarrow{h^w_{ijm}} = \overleftarrow{GRU}\left(x_{ijm}, \overleftarrow{h_{ij(m-1)}}\right), m \in \{M, ..., 1\} \tag{5.2}$$

The combination of the hidden state that is obtained from the forward GRU and the backward GRU $\overrightarrow{h^w_{ijm}}$ and $\overleftarrow{h^w_{ijm}}$ is represented as $h^w_{ijm} = \left[\overrightarrow{h^w_{ijm}} \oplus \overleftarrow{h^w_{ijm}}\right]$, which carries the complete tweet information centred around x_{ijm}. We explain the attention mechanism which involves introducing a trainable vector u_{ijm} for all words in order to capture global words. The annotations h^w_{ijm} form the foundation for attention which begins with another hidden layer. The model learns and randomizes biases (b_w) and weights (W_w) through training. The vector u_{ijm} for the annotations is represented as follows:

$$u_{ijm} = tanh(W_w h^w_{itm} + b_w) \tag{5.3}$$

The product $u_{ijm}u_w$ (u_w is randomly initialized) is expected to signal the importance of the m word and normalized to an importance weight per word α_{ijm} by a softmax function:

$$\alpha_{ijm} = \frac{\exp(u_{ijm}u_w)}{\sum_m \exp(u_{ijm}u_w)} \tag{5.4}$$

Finally, a weighted sum of word representations concatenated with the annotations previously determined called the tweet vector t_{ij}, where α_{ijm} indicating importance weight per word:

$$t_{ij} = \sum_m \alpha_{ijm} h^w_{ijm} \tag{5.5}$$

Tweet encoding: To learn the tweet representations h^t_{ij} from a pre-trained tweet vector t_{ij}, we capture the information of context at the tweet level. Similar to the word encoder component, the tweet encoder employs the same BiGRU architecture. Hence the combination of the hidden state that is obtained from the forward GRU and the backward GRU $\overrightarrow{h^t_{ij}}$ and $\overleftarrow{h^t_{ij}}$ is represented as $h^t_{ij} = \left[\overrightarrow{h^t_{ij}} \oplus \overleftarrow{h^t_{ij}}\right]$ which captures the coherence of a tweet concerning its neighbouring tweets in both directions. In order to measure the significance of the tweets, we once more employ the attention mechanism and introduce a context vector at the tweet level u_t. The results are computed as follows:

$$u_{ij} = tanh(W_t h^t_{ij} + b_t) \tag{5.6}$$

$$\alpha_{ij} = \frac{\exp(u_{ij} u_t)}{\sum_j \exp(u_{ij} u_t)} \tag{5.7}$$

$$c_i = \sum_j \alpha_{ij} h^t_{ij} \tag{5.8}$$

Where c_i is the vector that summarizes all the information of tweets in cluster i, which encode the word context and sentence context, respectively, and are learned jointly with the rest of the parameters.

Encoding clusters: After clustering the user tweets, our next goal is to model these clusters to obtain the cluster encoding that will play a key role in the predictive performance and in learning reliable explanations. Inspired by (210), we have developed a new model Hierarchical Attention Based Clustering Network (HACN) to learn user tweets representation at the cluster level. This model is depicted in Figure 5.4. We consider a dataset $D = [1,2,...,U]$ made of U users, and that a user has L tweets. The clustering method groups the user's tweets into E clusters $\{c_i\}_{i=1}^{E}$, and each cluster c_i contains J tweets $\{t_{ij}\}_{j=1}^{J}$. Each cluster is, thus, represented by its tweets and each

tweet is represented by the sequence of d-dimensional embeddings of their words $c_i = \{w_{i11}, w_{i12}, ..., w_{ijm}\}$, with $m \in [1, M]$ and $j \in [1, J]$, respectfully, represents the number of words in a tweet and the number of tweets in the i^{th} cluster.

Cluster encoding: We aim to locate tweets from users that can shed light on the reason for their depression and also help detect depression as they provide clear explainability. Since not all tweets from a user may have the same significance in determining if the user is depressed or not, we use an attention mechanism to determine the relevance of tweets in relation to depression and assign attention weights to them, resulting in more accurate and explainable predictions. We'll gather the relevant tweets in the formed vector $\hat{t}$ by implementing a tweet-level attention layer. The product $u_i u_s$ is intended to indicate the significance of the i tweet and is normalized to a per-tweet importance weight α_i. s_i is a vector that summarizes all the information in a user's tweets:

$$s_i = \sum_t \alpha_i h_i^t \tag{5.9}$$

and we summarize the equations below:

$$\overrightarrow{h_i} = \overrightarrow{GRU}\left(v_i, \overrightarrow{h_{i-1}}\right) \tag{5.10}$$

$$\overleftarrow{h_i} = \overleftarrow{GRU}\left(u_i, \overleftarrow{h_{i-1}}\right) \tag{5.11}$$

$$h_i = \left[\overrightarrow{h}_i \oplus \overleftarrow{h}_i\right] \tag{5.12}$$

$$u_i = tanh(W_s h_i + b_s) \tag{5.13}$$

$$\alpha_i = \frac{\exp(u_i u_s)}{\sum_t \exp(u_i u_s)} \tag{5.14}$$

$$\bar{s_i} = \sum_i a_i h_i \tag{5.15}$$

5.3.3 Hierarchical Attention Network for User Tweets

Word encoding: Similar to the word encoder in the cluster tweets encoding above, we represent each word as the input layer a fixed-size vector from pre-trained word embeddings. Each word of a user post (tweet) will convert to a word embedding x_{nq} utilizing GloVe (134). The combination of the hidden state that is obtained from the forward GRU and the backward GRU $\overrightarrow{h^w_{nq}}$ and $\overleftarrow{h^w_{nq}}$ is represented as $h^w_{nq} = \left[\overrightarrow{h^w_{nq}} \oplus \overleftarrow{h^w_{nq}}\right]$ which carries the complete tweet information centred around x_{nq}. We describe the attention mechanism where a weighted sum of word representations concatenated called the tweet vector b_n, where α_{nq} indicating importance weight per word:

$$b_n = \sum_q \alpha_{nq} h^w_{nq} \tag{5.16}$$

Tweet encoding: Given that (156) got all user tweets that were posted within a month of the anchor tweet, and as per in previous studies (178), people should be observed for a period of time; hence, it would be essential to explore the hierarchical user tweets its order during this month since these sequential tweets could contain valuable semantic information and have more potential to boost the depression detection when concatenated it with a cluster tweets. A user's posts contain language cues that differ in word and tweet level, as a result, the hierarchical attention network (210) can be more accurate to learn user tweets representation for this component as depicted in Figure 5.3. We now explain how to encode all user tweets to learn the latent representations.

Suppose the input which resembles all user tweet U be represented as $\{b_n\}^L_{n=1}$ where L is the total number of a user tweets and each tweet $b_n = \{w_{n1}, w_{n2}, ..., w_{nq}\}$ with $q \in [1, Q]$ represent, the maximum number of words in a user tweet. We now demonstrate how to encode all user tweets to learn the latent representations.

In order to learn the tweet representations h^b_n from a learned tweet vector b_n, we capture the information of context at the tweet level. Similar to the word encoder component, the tweet encoder employs the same BiGRU architecture. Hence the combination of the hidden state that is obtained from the forward GRU and the backward GRU $\overrightarrow{h^b_n}$ and $\overleftarrow{h^b_n}$ is represented as $h^b_n = \left[\overrightarrow{h^b_n} \oplus \overleftarrow{h^b_n}\right]$, which captures the coherence of a tweet concerning its neighbouring tweets in both directions. In order to measure the significance of the tweets, we once more employ the attention where a weighted sum of word representations concatenated called the tweet vector $\hat{b}_n$, where α_n indicating importance weight per word:

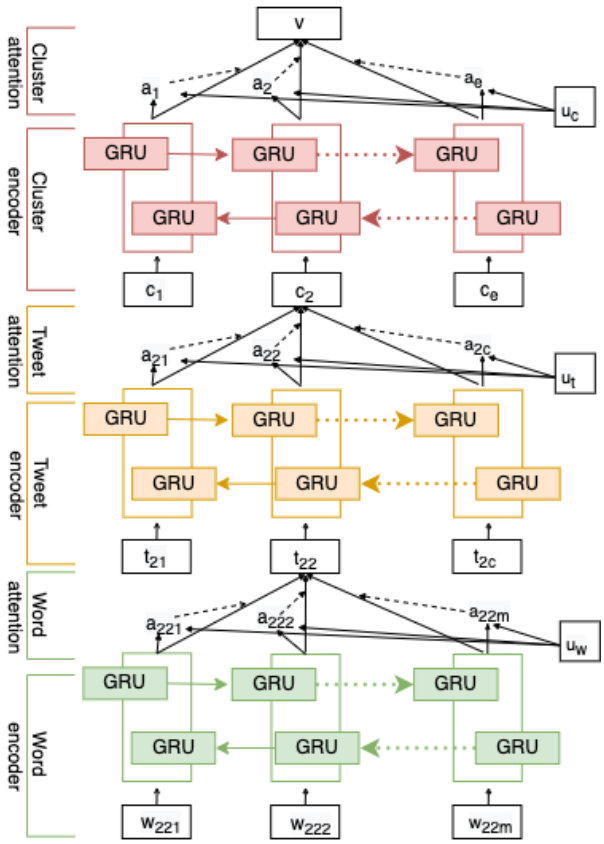


Figure 5.4: An illustration of Hierarchical Attention Network for Clustering of Tweets

$$\hat{b}_n = \sum_n \alpha_n h_n^w \tag{5.17}$$

$$p_i = f\left(b + \sum_{i=1}^{M} W_i m_i\right) \tag{5.18}$$

where f represents a nonlinear function, and the output of this function, p_i, is a high-level representation that captures the behavioural semantic information. This representation is crucial in the diagnosis of depression.

5.3.4 Prediction and Narrative Explainability

Modelling predictions: It is necessary to determine whether the user is suffering from depression or not. We have previously described how we process user behaviour features (p) and how we analyze user tweets by modeling the hierarchical structure from the

word level and tweet level (s). Subsequently, we integrate both components to create a feature matrix of user behavior features and user tweets to use in our classification task as follows:

$$p = p_1, p_2, ..., p_M \in \mathbb{R}^{1d \times M} \tag{5.19}$$

$$s = s_1, s_2, ..., s_n \in \mathbb{R}^{2d \times n} \tag{5.20}$$

We further unify these components together, which is denoted as $[p, s]$. The output of such a network is typically fed to a sigmoid layer for classification as follows:

$$\hat{y} = sigmoid(b_f + [p, s]W_f) \tag{5.21}$$

The predicted probability vector, $\hat{y}$, is composed of $\hat{y_0}$ and $\hat{y_1}$, which represent the predicted probability of the label being 0 (not depressed) and 1 (depressed user) respectively. Our objective is to minimize the cross-entropy error for each user with the true label, y:

$$\text{Loss} = -\sum_i y_i \cdot \log \hat{y}_i$$

where $\hat{y}_i$ is the predicted probability and y_i is the ground truth label (either depression or non-depression) user.

Modelling narrative and explanations: Our goal is to identify a specific narrative from a user that explains why they may be suffering from depression. By providing a clear explanation, it can also assist in detecting depression. The hierarchical attention mechanism we described above for cluster encoding is an effective method for assigning high weights to specific narrative elements. Additionally, the level of explainability of the user's topic is learned through attention weights. Since tweets in a cluster are assigned different weights based on the attention map, it demonstrates that our model can extract important and long-term contextual information from the cluster. In general, the attention map of our model can identify the most significant theme that relates to a depressed user and their corresponding group of tweets. Therefore, clusters with high attention weight are crucial and likely to provide an explanation for a user's depression.

Table 5.1: Summary of labelled data used to train NarrationDep model

Description	Depressed	Non-Depressed
Numer of users	2K	2K
Numer of tweets	0.5M	1M
Numer of clusters	33K	46K

5.4 Experiments and Results

This section presents an experimental assessment to compare **NarrationDep**'s performance with different strong comparative methods. We have used qualitative and quantitative techniques to evaluate our model and compare it with different strong comparative models.

5.4.0.1 Comparative methods

In Zogan et al, (224), the authors summarized user tweets utilizing a new summarization framework that is an interplay between extractive and abstractive summarization to generate a shorter representation of user historical tweets. As a result, this technique helps to reduce the influence of redundant content. They have already compared their method against several models that include convolutional neural networks with attention (CNN-Att) and Bidirectional Gated Recurrent Neural Networks with Attention (BiGRU-Att). Three pre-trained transformer models have also been investigated in the same paper which are XLNet (209), BERT (41) and RoBERTa (96). We have compared our model with the methods used in that work.

For all user tweets, and for the purpose of classifying user tweets, we employed the **BiGRU** (26) with the attention method that we employed to get user tweet representations. To represent user tweets and capture the semantics of various convolutional window sizes for depression detection, we used **CNN** (77) with an attention mechanism. Moreover, to identify depression in user postings, a hierarchical attention neural network architecture (210) is deployed. The network encrypts the first user postings by paying attention to both the words in each tweet and the tweets themselves. Finally, HCN is, similar to HAN, however instead of utilizing (GRU), hierarchical convolutional networks (HCN) rely on architectures based on convolutional neural networks.

5.4.1 Datasets

To evaluate the effectiveness of the models, we conduct our experiments on Shen et al., (156) pre-COVID dataset, as shown in Table 5.1, which contains users and their posts on Twitter. Each user is labelled either depressed or non-depressed. The authors labelled users as depressed if they found a user's tweet that contained a specific pattern. Between 2009 and 2016, they constructed depressed users based on the content of their posts. There are around 2000 depressed users and around 400K tweets in all. The tweets for non-depressed were collected in December 2016, including 2000 users and over a million tweets. For preprocessing, we remove users who have less than ten tweets and for evaluation, we randomly split the dataset into training and test sets with a ratio of 80:20 with 5-fold cross-validation.

Table 5.2: Performance comparison on (156). NarrationDep vs all user tweets summarization

Model	Precision	Recall	F1-Score	Accuracy
XLNet (base)	0.868	0.843	0.848	0.835
BERT (base)	0.766	0.762	0.786	0.764
RoBERTa (base)	0.817	0.804	0.786	0.806
BiGRU-Att	**0.941**	0.731	0.823	0.836
CNN-Att	0.861	0.843	0.835	0.837
CNN_BiGRU-Att	0.836	0.829	0.824	0.824
NarationDep	0.884	**0.878**	**0.878**	**0.879**

Table 5.3: Performance comparison on (156). NarrationDep vs All user tweets training data

Model	Precision	Recall	F1-Score	Accuracy
BiGRU	0.766	0.762	0.786	0.764
CNN	0.817	0.804	0.786	0.806
HAN	0.87	0.844	0.856	0.835
HCN	0.853	0.852	0.852	0.852
NarationDep	**0.884**	**0.878**	**0.878**	**0.879**

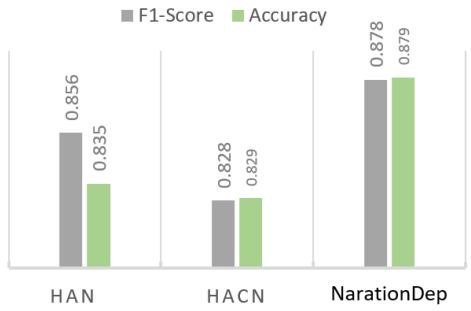


Figure 5.5: Impact analysis of all user contents model (HAN), clustering tweets model (HACN), and our model (NarationDep) for depression detection.

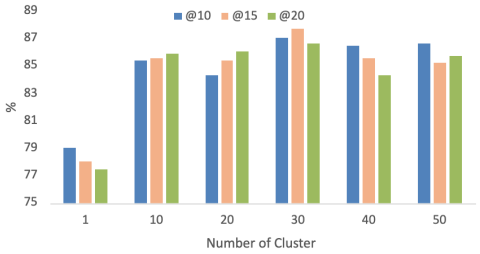


Figure 5.6: Effectiveness of our model (NarationDep) with different numbers of clusters

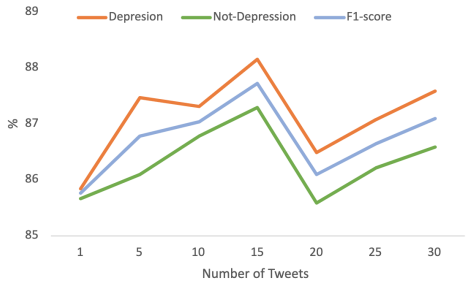


Figure 5.7: Effectiveness of our model (NarationDep) using F1 score for each class (Depression and non-Depression) with different number of tweets

5.4.2 Evaluation Metrics and Settings

We employ the standard metrics popularly used in information retrieval which are accuracy, F1-score, and precision to compare the classification performance. These metrics are widely used in previous works for depression detection (156; 227; 224; 223).

We performed the experiments in Python 3.6.3 environment and Tensorflow 2.1.0. Word embeddings is initialized using Glove (134). The dimension of word embedding is 100, and the dropout rate is set to 0.5. We used the Adam optimization algorithm for both HCN and HCN+ with default value learning rate (lr) = 0.001. We train HCN+ for 20 epochs on all the data with a batch size of 32.

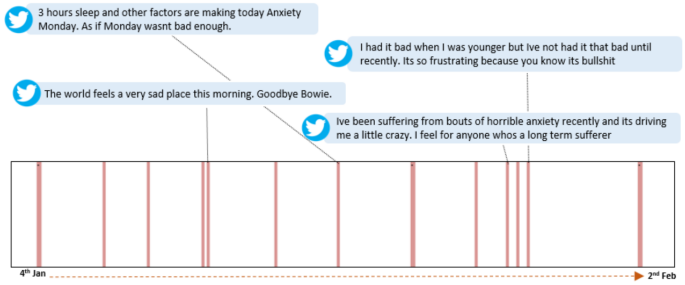


Figure 5.8: Narrative tweets captured by ***NarationDep***

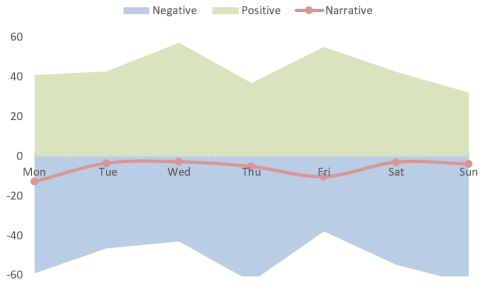


Figure 5.9: *Analyzing narrative for a user per week.*

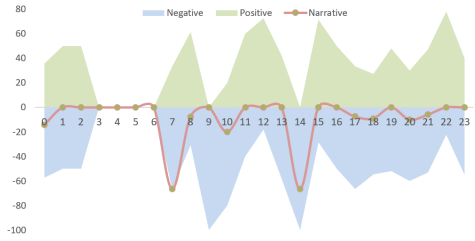


Figure 5.10: *Analyzing narrative for a user during the 24 hours of the day.*

5.4.3 Quantitative Results

We now report the quantitative results obtained from different models. Evaluation results for different competing methods are presented, where the best results for the best model are highlighted in bold in Table 7.2. The first part of the table shows the effectiveness when using summarization sequences of user posts to detect depression; the performance compared using some models that achieved a new state-of-the-art result on many NLP tasks, such as text classification. We see that CNN_BiGRU-At outperforms other models with F1-score and recall; however, XLNet and RoBERTa perform best among all the different models with accuracy and precision, respectively.

The second part of the results is presented in Table 5.3, where we show the results of using all user tweets; we observed that all the hierarchical text classification models (HAN and HCN) outperform other neural network-based methods, such as BiGRU and CNN. Furthermore, we observed that our model ***NarationDep*** outperforms other models in terms of Accuracy, F1/-Score and Recall. Comparing our model with HAN, ***NarationDep*** boosts about 4.4%, 2.2%, 3.4% and 1.4% in terms of Accuracy, F1-score, Recall and Precision.

Generally, our models based on the hierarchical network can consistently outperform

other methods in terms of accuracy, F1 Score and recall on both training data (tweets summarization and all user tweets). Our models based on hierarchical networks achieve a relative improvement of 3.2% for ***NarationDep***, compared against the best results (XLNet) in terms of Accuracy.

To evaluate the performance of our model, we have compared the effectiveness of each of its two components. We test the model by feeding it with each component separately and comparing its performance. First, we test the model using only the HACN (Hierarchical Attention Based Clustering Network) for clustering tweets. As seen in Figure 5.5, the model performs less optimally when we used **HACN** alone. In contrast, when we use only HAN with word embedding attributes, the model performs better. This indicates that extracting semantic information from user tweets is essential for depression detection. Therefore, we can see that the performance of the *NarationDep* model improves when HACN and HAN are combined and outperforms when using each attribute alone. One of the crucial parameters in *NarationDep* is the number of clusters for each user in HACN; we found that *NarationDep* achieves optimal performance when using 30 clusters as the maximum number of clusters. The performance of the model with respect to the number of clusters is shown in Figure 5.6. In Figure 5.7, we present the impact of the number of tweets.

5.4.4 Qualitative Study

We conduct a qualitative study to demonstrate the effectiveness of our model on explainability. We primarily exploit the attention weights that are learned by our model to find out whether tweets that convey some of a narrative are assigned a high attention weight or not. Our motivation is that those tweets that are shared by a user that convey some linked narrative must be assigned a high attention weight by our model.

We depict the attention map for a randomly selected depressed user in Figure 5.8 to show the narrative tweets that **NarationDep** captured. We emphasize the model's significance associated with the explainability of depression detection. The bar, in the figure, represents all user tweets, and since **NarationDep** can indicate the more critical cluster with importance by attention weights; we can argue that all the tweets of this cluster depict a user narrative. The weight of the narrative tweets is represented in red colour in the figure. Different tweets have varying weights within each cluster as determined by the attention map.

What we learn from this qualitative study is that **NarationDep** can extract vital and in-depth contextual information from a cluster. In general, the attention map of

NarationDep can identify the most relevant cluster for determining a depressed user, as shown by the red lines in the bar chart and their corresponding tweets, while clusters containing tweets that do not contribute to classifying a depressed user have low attention weights and will be ignored. The figure shows that the attention map assigns higher weights to tweets that provide an explanation for depression; for example, tweet *"Ive been suffering from bouts of horrible anxiety ... "* attained the highest attention score among all the tweets from the user. Additionally, **NarationDep** assigns higher weights to tweets that provide explanations compared to interfering or unrelated tweets, allowing for the selection of more relevant tweets and making it a more useful feature for identifying depressed users.

In Figures 5.9 and 5.10, we analyse the temporal dynamics associated with the user. We analyse the user narrative in a week followed by modelling their daily narrative. The lines in the two figures depict how the user narratives move between positive and negative narratives.

5.5 Summary

Understanding a user's social media narrative is an individual story that is broken up into various posts across different chronologies and platforms such as Facebook, Twitter, LinkedIn, and Instagram. Even though the individual posts are short, when combined, they create a more comprehensive story with some key themes and continuity. It can be challenging yet useful to automatically detect mental health issues with this type of data because it is short, infrequent, and sometimes poorly worded. However, efforts can be made to develop computational models to automatically identify patterns in user-generated content and extract key information for a narrative explanation of a series of events which would be significant if we compare people with a mental disorder.

In this work, we studied a novel problem of modelling narrative elements in social media to analyze how our posts can be used to understand a user narrative. We develop a novel model which is a user hybrid classification model ***NarationDep*** to automatically detect depressed users based on a hierarchical attention network, which exploits data from Twitter. ***NarationDep*** has a component called HACN, which considers the hierarchical structure of user cluster tweets (cluster, tweets and words) and contains an attention mechanism that can find the most crucial cluster that represents the narrative explanation in a user document. We expand user input by using two attributes and read the input in different ways in parallel using HAN and HACN to extract different

features from the same input to boost our model performance. The results showed that our model ***NarationDep*** with two attributes outperform strong comparative models and effectively detect depressed users. There are a number of directions that we could take after this work. In the future, we will develop a multi-modal model that will exploit key features from videos and images because they contain complementary information. While challenging, we can develop the method by exploiting visual-language models such as DALL-E (138). Another direction is automatically predicting relevant diagnoses to the mental health problem that has received little attention. This is an important problem to address because if the model could reliably predict whether a user is depressed or not, the model must be reliable enough to also propose certain diagnoses to the issue. As a result, it will make the process efficient and more mental health disorders could be addressed in less time. While human-in-the-loop is important, over time, by exploiting models in reinforcement learning, the models could be improved.

CHAPTER 6

COMMUNITY DEPRESSION DYNAMICS DETECTION DURING COVID-19

6.1 Background and Motivation

The outbreak of the novel Coronavirus Infectious Disease 2019 (COVID-19) has caused an unprecedented impact on people's daily lives around the world (204). People's lives are at risk because the virus can easily spread from person to person (169) either by coming in close contact with the infected person or sometimes may even spread through community transmission[1], which then becomes extremely challenging to contain. The infection has now rapidly spread across the world and there have been more than 10.3 million confirmed cases and more than 505,000 people have died because of the infection until 30 June 2020[2]. Almost every country in the world is battling against COVID-19 to prevent it from spreading as much as possible. While some countries such as New Zealand has been very successful in containing the spread, others such as Brazil and India have not. As a result, this outbreak has caused immense distress among individuals either through infection or through increased mental health issues, such as depression, stress, worry, fear, disgust, sadness, anxiety (a fear for one's health, and a fear of infecting others), and perceived stigmatisation (112; 13; 147). These mental health issues can even occur in people, not at high risk of getting infected. There could be even several

[1] https://www.who.int/publications/i/item/preparing-for-large-scale-community-transmission-of-covid-19

[2] https://coronavirus.jhu.edu/

people who are exposed to the virus may be unfamiliar with it as they may not follow the news, or are completely disconnected with the general population. (112).

Consider depression as an example, which is the most common mental health issue among other mental health issues according to the World Health Organisation (WHO), with more than 264 million people suffering from the depression worldwide (198). Australia is one of the top countries where mental health disorders have high proportions over the total disease burden (see Fig. 6.1). Depression can cause severe emotional, behavioural and physical health problems. For example, people with depression may experience symptoms amounting to their inability to focus on anything, they constantly go through the feeling of guilt and irritation, they suffer from low self-worth, and experience sleep problems. Depression can, therefore, cause serious consequences, at both personal and social costs (187).

A person can experience several complications as a result of depression. Complications linked to mental health especially depression include: unhappiness and decreased enjoyment of life, family conflicts, relationship difficulties, social isolation, problems with tobacco, alcohol and other drugs, self-harm and harm to others (such as suicide or homicide), weakened immune system. Furthermore, the consequence of depression goes beyond functioning and quality of life and extends to somatic health. Depression has been shown to subsequently increase the risk of, for example, cardiovascular, stroke, diabetes and obesity morbidity (135). However, the past epidemics can suggest some cues of what to look out for after COVID-19 in the next few months and years. For example, when patients with SARS and MERS were assessed a few months later, 14.9% had depression and 14.8% had an anxiety disorder (147).

Meanwhile, to reduce the risk of the virus spreading among people and communities, different countries have taken strict measures such as locking down the whole city and practising rigorous social-distancing among people. For example, countries such as China, Italy, Spain, and Australia are fighting the COVID-19 pandemic through nation-wide lockdown or by cordoning off the areas that were suspected of having risks of community spread throughout the pandemic, expecting to “flatten the curve”. However, the long-term social activity restriction policies adopted during the pandemic period may further amplify the mental health issues of people. Therefore, it is important to examine people’s mental health states as a result of COVID-19 and related policies, which can help governments and related agencies to take appropriate measures more objectively if necessary.

[3] `https://ourworldindata.org/`

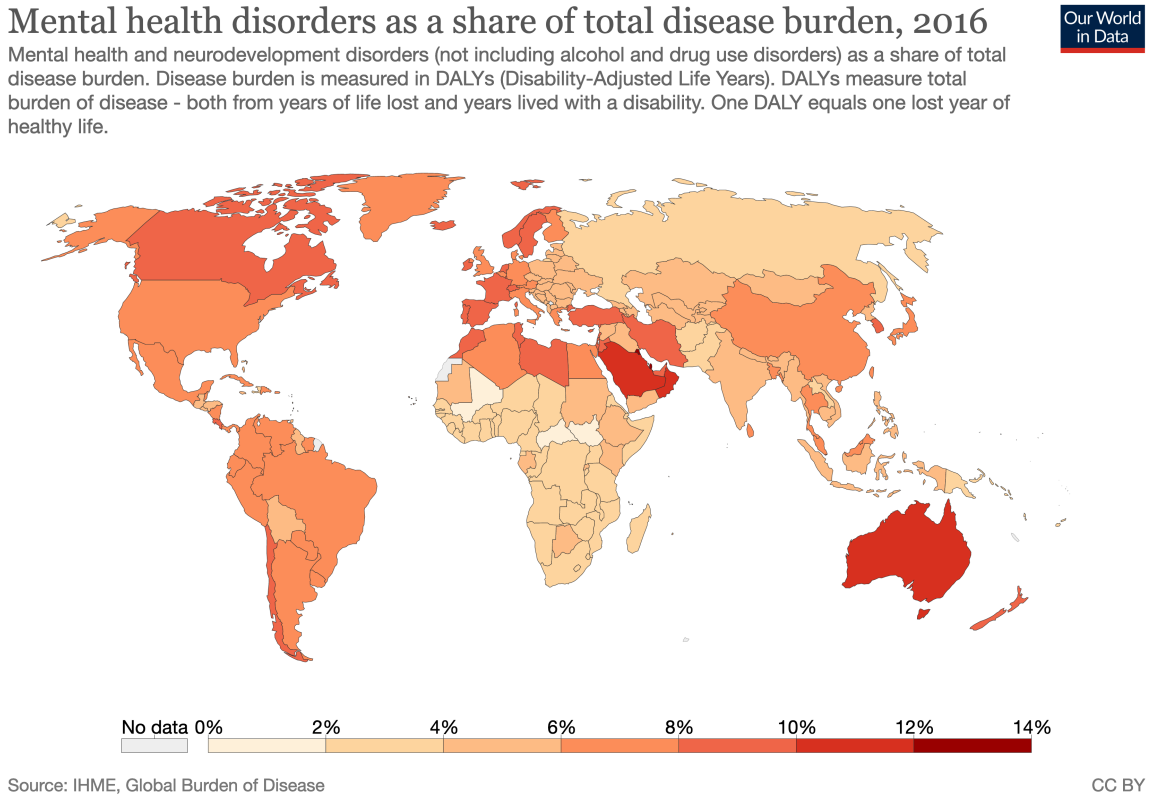


Figure 6.1: The world mental health disorders in 2016[3].

On the other hand, we have witnessed increased usage of online social media such as Twitter during the lockdown[4]. For instance, 40% consumers have spent longer than usual on messaging services and social media during the lockdown. It is mainly because people are eager to publicly express their feelings online given an unprecedented time that they are going through both physically and emotionally. The social media platforms represent a relatively real-time large-scale snapshot of activities, thoughts, and feelings of people's daily lives and thereby reflect their emotional well-being. Every tweet represents a signal of the users' state of mind and state of being at that moment (55; 215). Aggregation of such digital traces may make it possible to monitor health behaviours at a large-scale, which has become a new growing area of interest in public health and health care research (214; 72; 151).

Since social media is social by its nature, and social patterns can be consequently found in Twitter feeds, for instance, thereby revealing key aspects of mental and emotional disorders (33). As a result, Twitter recently has been increasingly used as a

[4]https://www.statista.com/statistics/1106498/home-media-consumption-coronavirus-worldwide-by-country/

viable approach to detect mental disorders of depression in different regions of the world (142; 6; 87; 141; 104). For example, the research found that the depressed users were less active in posting tweets, doing it more frequently between 23:00 and 6:00. The use of vocabularies could also be an indicator of depression in Twitter, for example, it was found that the use of verbs was more common by depressed users, and the first-person singular pronoun was by far the most used by depressed users (87). Hence, many research work has been done to extract features like user's social activity behaviours, user profiles, texts from their social media posts for depression detection using machine learning techniques (36; 178; 208; 71; 156). For example, De Choudhury (36) et al., proposed to predict depression for social media users based on Twitter using support vector machine (SVM) for prediction based on manually labelled training data.

The most recent work using Twitter to analyse mental health issues due to COVID-19 are (11; 13; 222). These work focus more on public sentiment analysis. Furthermore, little work such as Li et al., (89) classify each tweet into the emotions of anger, anticipation, disgust, fear, joy, sadness, surprise and trust. The two emotions of sadness and fear are more related to severe negative sentiments like depression due to COVID-19. However, little work is done to detect depression dynamics at the state level or even more granular level such as suburb level. Such granular level analysis of depression dynamics not only can help authorities such as governmental departments to take corresponding actions more objectively in specific regions if necessary but also allows users to perceive the dynamics of depression over the time to learn the effectiveness of policies implemented by the government or negative effects of any big events. The questions for which we wish to find the answers are:

- How people's depression is affected by COVID-19 in the time dimension in the state level?
- How people's depression is affected by COVID-19 in the time dimension in local government areas?
- Can we detect the effects of policies/measures implemented by the government during the pandemic on depression?
- Can we detect the effects of big events on depression during the pandemic?
- How effective is the model in detecting people's depression dynamics?

This chapter aims to examine community depression dynamics due to COVID-19 pandemic in Australia. A new approach based on multi-modal features from tweets and

term frequency-inverse document frequency (TF-IDF) is proposed to build a depression classification model. Multi-modal features aim to capture depression cues from emotion, topic and domain-specific perspectives. TF-IDF is simple, scalable, and has proven to be very effective to model a wide-range of lexical datasets including large and small datasets. In contrast, recent computationally demanding frameworks such as deep learning, which are usually generated rely upon large datasets because that helps give them faithful co-occurrence statistics might not be obtained from sparse and short texts from user-generated content such as tweets. Our approach uses the TF-IDF model can generalise well in various situations where both small and large datasets can be used leading to a reliable and scalable model. After building the model for depression classification, Twitter data in the state of New South Wales in Australia are then collected and input to our depression classification model to extract depression polarities which may be affected by COVID-19 and related events during the COVID-19 period. The contributions of this chapter primarily include:

- Novel multi-modal features including emotion, topic and domain-specific features are extracted to describe depression in a comprehensive approach;
- A faithful depression classification model based on TF-IDF, which is simple and scalable in generalisation, is proposed to detect depressions in short texts such as tweets;
- Instead of the depression examination of the whole country, a fine-grained analysis of depression in local government areas of a state in Australia is investigated;
- The links between the community depression and measures implemented by the government during the COVID-19 pandemic are examined.

To the best of our knowledge, this study is the first work to investigate the community depression dynamics across the COVID-19 pandemic, as well as to conduct the fine-grained analysis of depression and demonstrate the links of depression dynamics with measures implemented by the government and the COVID-19 itself during the pandemic.

The remainder of the paper is organized as follows. We firstly review the related work in Section II and introduce the collected real-world dataset in Section III. After that, we demonstrate our proposed method for COVID-19 depression analysis in section IV and present the experiments and verify the performance of the proposed model in Section V. The proposed novel model is then used to detect community depressions in

New South Wales in Australia in Section VI. Finally, this work is concluded with an outlook on future work in Section VII.

6.2 Preliminary on Community Depression Dynamics Detection During COVID-19

In this section, we review the related work for depression detection. We also highlight how our work differs from these existing approaches.

6.2.1 Machine learning based depression detection

Social media has long been used as the data source for depression detection due to the largely available user-generated text data (207; 84; 228). The shared text data and the social behaviour of the social network users are assumed to contain clues for identifying depressed users. To find the depression pattern for social media users, many works have been done to adopt traditional machine learning models such as Support Vector Machine (SVM) and J48 for depression classification and detection based on different feature engineering techniques. For example, Wang et al. (197) have proposed a binary depression detection model for Chinese Sina micro-blogs to classify the posts as depressed or non-depressed. Based on the features extracted from the content of the micro-blog such as the sentiment polarity of sub-sentences, the users' online interactions with others and the user behaviours, they have trained J48 tree, Bayes network, and rule-based decision table as classifiers for depression detection.

De Choudhury (36) et al. have investigated to predict depression for social media users based on Twitter and found that social media contain meaningful indicators for predicting the onset of depressions among individual users. To get the ground truth of users' suffered depression history, De Choudhury et al. adopted the crowdsourcing to collect Twitter users who have been diagnosed with clinical (Major Depressive Disorder) MDD based on the CES-D2 (Center for Epidemiologic Studies Depression Scale) screening test. Then, to link the depression symptoms with the social media data, they extracted several measures such as user engagement and emotion, egocentric social graph, linguistic style, depressive language user, and the mentions of antidepressant medications from users' social media history for one year. Finally, they trained SVM as the depression classifier base on the ground truth and the extracted features for the tested Twitter users with prediction accuracy around 70%. Similarly, Tsugawa et al. (178) have investigated in

recognizing depression from the Twitter activities. To get the ground truth data of users' depression degree record, they chose the web-based questionnaire for facts collection. Then, similar features such as topic features extracted using topic modelling like Latent Dirichlet Allocation (LD), polarities of words and tweet frequency are extracted from the Twitter users' activity histories for training an SVM classifier.

Later, Yang et al. (208) have proposed to analyse the spatial patterns of depressed Twitter users based on Geographic Information Systems (GIS) technologies. They firstly adopted Non-negative Matrix Factorization (NMF) to identify the depressed tweets from online users. Then, they exploited the geo-tagged information as the indicator of users' geographical location and analyzed the spatial patterns of depressed users. Shen et al. (156) have made efforts to explore comprehensive multi-modal features for depression detection. They adopted six groups of discriminant depression-oriented features extracted from users' social network, profile, visual content, tweets' emotions, tweets' topics and domain-specific knowledge are used as representations for Twitter users. Then, a dictionary learning model is used to learn from the features and capture the jointly cross-modality relatedness for a fused sparse representation and train a classifier to predict the depressed users.

Different from these work in feature extraction which may be either impractical like using crowdsourcing to manually label each user or too customized like extracting different kinds of social behaviours of users, we propose to combine the robust and simple text feature representation method: Term Frequency-Inverse Document Frequency (TF-IDF) with other multi-modal features such as topics and emotions to represent the text data.

6.2.2 Deep learning based depression detection

More recently, the rapidly developed deep learning techniques have also been used for depression detection. For example, Shen et al. (158) have proposed a cross-domain deep neural network with feature transformation and combination strategy for transfer learning of depressive features from different domains. They have argued that the two major challenges regarding cross-domain learning are defined as isomerism and divergence and proposed DNN-FATC which includes Feature Normalization & Alignment (FNA) and Divergent Feature Conversion (DFC) to better transfer learning. Orabi et al. (71) have proposed to explore the word embedding techniques where they used pre-trained Skip-Gram (SG) and Continuous Bag-of-Words (CBOW) models currently implemented in the word2vec (108) package for better textual feature extraction to train a neural

network-based classifier. We have also seen researchers using deep reinforcement learning to depression detection on Twitter. For instance, Gui et al. (58) have proposed a cooperative multi-agent model to jointly learn the textual and visual information for accurate depression inference. In their proposed method, the text feature is extracted using a Gated Recurrent Unit (GRU) and the visual feature is extracted using Convolutional Neural Networks (CNN). The selection for useful features from GRU and CNN is designed as policy gradient agents and trained by a centralized critic that implements difference rewards.

Even though deep learning has become the dominated method for many classifications or prediction tasks, it cannot guarantee that deep learning is feasible on any tasks. For example, (119) compared a bunch of methods including different conventional methods and deep learning methods for information extraction tasks from a text corpus. We are not surprised to find that deep learning did not outperform the conventional methods as expected. Furthermore, deep learning techniques are usually dependent on large datasets with faithful co-occurrence statistics which might not be obtained from sparse and short texts especially online user-generated content such as tweets which is characterised by relatively limited labelled text in the depression study. While TF-IDF is simple and scalable, it has proven to model a wide-range of datasets including large and small datasets. Therefore, in our study, we propose to adopt the traditional classification methods with TF-IDF for depression detection to obtain a faithful model with a strong generalisation ability in various situations of small and large datasets. Most importantly, even non-experts such as those from the government and NGOs could easily understand the intricacies of the model and apply our model on their data to detect community dynamics in their region and make further decisions accordingly.

6.2.3 Depression detection due to COVID-19

The impact of COVID-19 on people's mental health has been recently reported in various research. For instance, Galea et al. (52) have pointed out that the mental health consequences due to this pandemic are applicable in both the short and long term. They have appealed to develop ways to intervene with the inevitability of loneliness and its consequence as people are physically and socially isolated. Huang et al. (69) have exploited a web-based cross-sectional survey based on the National Internet Survey on Emotional and Mental Health (NISEMH) to investigate people's mental health status in China by spreading the questionnaire on Wechat (a popular social media platform in China). To infer the depression symptoms for the anonymous participants, they have

adopted a predefined Epidemiology Scale to identify whether participants had depressive symptoms. Similarly, Ni et al. (121) have conducted a survey on the online platform to investigate the mental health of 1577 community-based adults and 214 health professionals during the epidemic and lockdown. The results show that around one-fifth of respondents reported probable anxiety and probable depression. These works are mainly based on questionnaires and pre-defined mental heath scale models for inference. In contrast, our proposed work relies on detecting depression from social media data automatically which shows advantages in monitoring a large number of people's mental health states.

6.3 Data

6.3.1 Study location

In this study, a case study for analysing depression dynamics during the COVID-19 pandemic in the state of New South Wales (NSW) in Australia is conducted. NSW has a large population of around 8.1 million people according to the census in September 2019 from Australian Bureau of Statistics[5]. The NSW's capital city Sydney is Australia's most populated city with a population of over 5.3 million people. The Local Government Areas (LGAs) are the third tier of government in the Australian state (the three tiers are federal, state, and local government).

6.3.2 Data collection

To analyse the dynamics of depression during the COVID-19 pandemic period at a fine-grained level, we collected tweets from Twitter users who live in different LGAs of NSW in Australia. The time span of the collected tweets is from 1 January 2020 to 22 May 2020 which covers dates that the first confirmed case of coronavirus was reported in NSW (25 January 2020) and the first time that the NSW premier announced the relaxing for the lockdown policy (10 March 2020). There are 128 LGAs in NSW. In this study, Twitter data were collected for each LGA separately so that the depression dynamics can be analysed and compared for LGAs. Twitter data were collected through the user timeline crawling API *user_timeline*[6] in Tweepy which is a python wrapper for the official

[5] `https://www.abs.gov.au/`

[6] `http://docs.tweepy.org/en/latest/api.html#api-reference`

Table 6.1: Summary of the collected Twitter dataset.

Description	Size
Total Twitter users	183,104
Average Twitter user per LGA	1,430.5
Average tweets per LGA	739,900.5
Total tweets	94,707,264

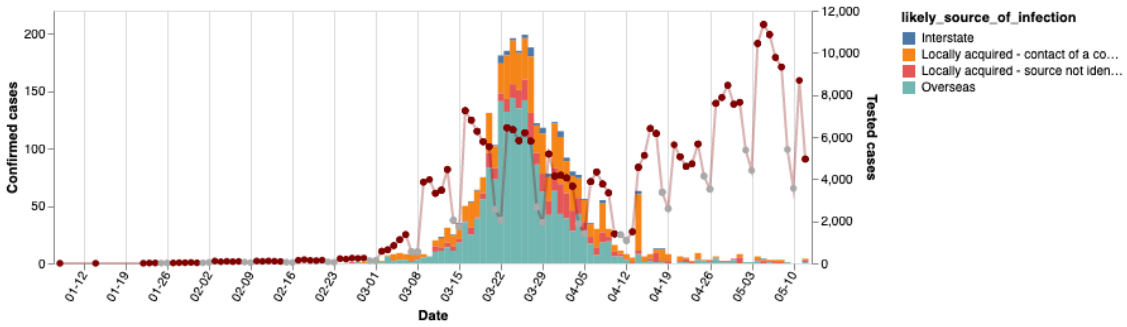


Figure 6.2: The tests and confirmed cases of COVID-19 in NSW until 22 May 2020.

Twitter API [7]. Table 6.1 shows the summary of the collected tweet dataset. In summary, 94,707,264 tweets were collected with averagely 739,901 tweets for each LGA during the study period. Datasets of COVID-19 tests and confirmed cases in NSW during the study period were collected from DATA.NSW [8].

Fig. 6.2 shows the overview of the number of tests (polylines with dots) and confirmed cases (bars) of COVID-19 in NSW over the study period. It demonstrates that there usually had test peaks at the beginning of each week and had fewer numbers at the weekend, which well aligns with the people's living habits in Australia. It shows that the outbreak peak of COVID-19 in NSW was on 26 March 2020 and tests were significantly increased after 13 April 2020. It also shows that most of the confirmed cases were originally related to overseas.

6.3.3 Dataset for depression model training

In order to detect depression at the tweet level, we created two labelled datasets for both depressed and non-depressed tweets:

- Positive tweets (depressed): we used a previous work dataset from (156). Shen et al. (156) published around 300K tweets with 1400 depressed users. In order to

[7] https://developer.twitter.com/en/docs/api-reference-index
[8] https://data.nsw.gov.au/

have a better performance, we increased the number of positive tweets by crawling additional 600K tweets from the Twitter keyword streaming API. We adopted the same keywords search that selected by (156) where users identified themselves as depressed, and we also used a regular expression to find positive tweets (e.g. I'm depressed AND suicide).

- Negative tweets (non-depressed): In order to balance the negative tweets with the positive tweets, we randomly selected 900K tweets which were not labelled as depressed from the collected tweets. Table 6.2 shows the summary of labelled data used to train the depression model.

Table 6.2: Summary of labelled data used to train depression model.

Description	**Size**
Depressed tweets	~ 900K
Non-Depressed tweets	~ 900K

After the collection of experimental data, features need to be extracted from social media text. However, because of the free-style nature of social media text, it is hard to extract useful features directly from raw text data and apply them to a classifier. The raw text data also affect the efficiency of extracting reliable features, and it makes it difficult to find word similarities and apply semantic analysis. Therefore, raw data must be pre-processed before feature extraction, in order to clean and filter data and ensure the data quality for feature extraction. Pre-processing may also involve other procedures such as text normalization.

Natural Language Processing (NLP) toolkit has been widely used for text pre-processing due to its high-quality processing capabilities such as processing sentimental analysis datasets (65). Natural Language Processing Toolkit (NLTK) library is considered as one of the most powerful NLP libraries in Python programming. NLTK contains packages that make data processing with human language easily and is used widely by various researchers for text data pre-processing. Therefore, before feeding our data to the model, we used NLTK to remove user mentions, URL links and punctuation from tweets. Furthermore, we removed common words from each tweet such as "the", "an", etc.). There are various reasons which have been mentioned in the literature where removing the stop words has had a positive impact on the model's quantitative performance, for instance, sometimes stop words deteriorate the classifications performance, sometimes they also have a huge impact on the model efficiency because these stop words increase

the parameter space of the model, among various other reasons. NLTK has a set of stop words which enable removing them from any text data easily. Finally, we stem tweets using NLTK using the Porter Stemmer.

6.4 Our Model

6.4.1 Proposed method

In this study, two sets of features are extracted from raw text and used to represent tweet. Since extracting features would be challenging due to the short length of the tweets and single tweet does not provide sufficient word occurrences, we, therefore, combine multi-modal feature with Term Frequency-Inverse Document Frequency (TF-IDF) feature to analyze depressed tweets. Our proposed framework is shown in Fig. 6.3.

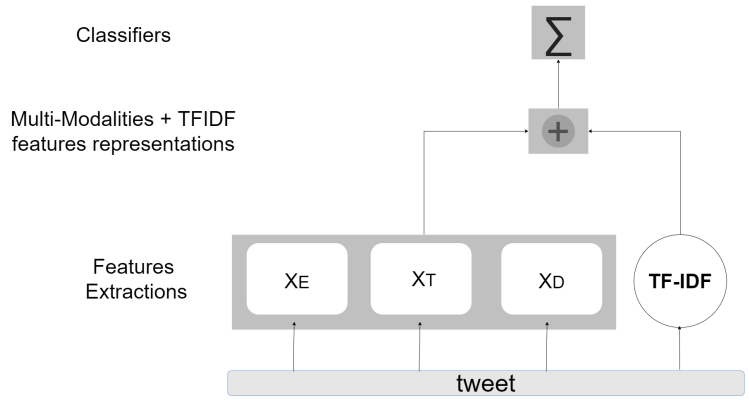


Figure 6.3: The proposed framework to detect depressed tweets during the COVID-19

6.4.2 Multi-modal features

User behaviours at the tweet level could be modelled by the linguistic features of tweets posted by the user. Inspired by (156), we defined a set of features consisting of three modalities. These three modalities of features are as follows:

- **Emotional features**: The emotion of depressed people is usually different from non-depressed people, which influences their posts on social media. In this work, we studied user positive and negative emoji in each tweet to represent emotional features. Furthermore, users in social media often use a lot of slang and short words,

which also convey positive and negative emotions (122). In this study, positive and negative emotion features are also extracted based on those slang and short words.

- **Topic-level features**: We adopted Latent Dirichlet Allocation (LDA) (15) to extract topic-level features since LDA is the most popular method used in the topic modelling to extract topics from text, which could be considered as a high-level latent structure of content. The idea of LDA is based on the assumption of a mixture of topics forms documents, each of which generates words based on their Dirichlet distribution of probability. Given the scope of the tweet content, we defined 25 latent topics in this study, which topic number is often adopted in other studies. We have also found that this number of topics gives satisfactory results in our experiments. we implemented LDA in Python with Scikit-Learn.

- **Domain specific features**: Diagnostic and Statistical Manual of Mental Disorders 4th Edition (also called DSM-IV) is a manual published by the American Psychiatric Association (APA) that includes almost all currently recognized mental health disorder symptoms (123). We, therefore, chose the DSM-IV Criteria for Major Depressive Disorder to describe keywords related to the nine depressive symptoms. Pre-trained word2vec (Gensim pre-trained model based on Wikipedia corpus) was used in this study to extend our keywords that are similar to these symptoms. We also extracted "Antidepressant" by creating a complete list of clinically approved prescription antidepressants in the world.

    1. Depressed mood.
    2. Loss of interest
    3. Weight or appetite change
    4. Sleep disturbance
    5. Psychomotor changes
    6. Fatigue or loss of energy
    7. Feel Worthlessness
    8. Reduced concentration
    9. Suicidal ideation

For a given sample of tweet, the multi-modal features are represented as $X_{t1}, X_{t2}, X_{t3}, \ldots, X_{tn}$, where $X_{ti} \in \mathbb{R}^{\mathbf{d}}$ is the d-dimensional feature for the i-th modality for each tweet and n is the size of the combined feature space, which is 21 in this study.

6.4.3 TF-IDF

We first review the definitions of Term Frequency and Inverse Document Frequency below.

Definition 1. Term Frequency (TF): consider the tweets in Bag-of-Word model (BoW), where each tweet is modeled as a sequence of words without order info. Apparently in BoW scheme, a word in a tweet with occurrence frequency of 10 is more important than a term with that frequency of 1, but not proportional to the frequency. Here we use the log term frequency ltf as defined by:

$$ltf_{(t,d)} = 1 + log(1 + tf_{(t,d)}) \tag{6.1}$$

where $tf_{(t,d)}$ represents occurrence number of the term t in a tweet d.

Definition 2. Inverse Document Frequency (IDF): It uses the frequency of the term in the whole collection for weighting the significance of the term in light of inverse effect. Therefore under this scheme, the IDF value of a rare word is higher, whereas lower for a frequent term, i.e. weighting more on the distinct words. The log IDF which measures the informativeness of a term is defined as:

$$idf_t = log_{10} \frac{N}{df_t} \tag{6.2}$$

where N represents the total number of tweets in the tweet corpus, and df_t the number of tweets containing the term t.

Therefore, TF-IDF is calculated by combining TF and IDF as represented in Eq.6.3.

$$tfidf_t = tf_t * idf_t \tag{6.3}$$

In order to extract relevant information from each tweet and reduce the amount of redundant data without losing important information, we combined multi-modalities with TF-IDF. TF-IDF is a numerical statistic metric to reflect the importance of a word in a list or corpus, which is widely studied in relevant work. Choudhury et al. (36) applied TF-IDF to words in Wikipedia to remove extremely frequent terms and then used the top words with high TF-IDF. The approach helped them to assess the frequency of uses of depression terms that appear on each user's Twitter posts on a given day. In (164) the authors have used TF-IDF in their model to compare the difference of performance by feeding TF-IDF into five machine learning models. It was found that all of them can achieve very good performance. However, one weakness of BoW is unable to consider the word position in the text, semantics, co-occurrences in different documents. Therefore, TF-IDF is only useful as a lexical level feature.

Table 6.3: The performance of tweet depression detection based on multi-modalties only.

Features	Method	Precision	Recall	F1-Score	Accuracy
Multi-Modal	LR	0.842	0.828	0.832	0.833
	LDA	0.843	0.816	0.820	0.824
	GNB	0.873	0.814	0.818	0.825

6.4.4 Modeling depression in tweets

Two labelled datasets as introduced in the previous section are used to train the depression classification model for tweets. Here we only use English tweets to train the model, and all non-English tweets are excluded. We also exclude any tweets with a length shorter than five words since these tweets could only introduce noise and influence the effectiveness of the model negatively. Three mainstream classification methods are used in this study to compare their performance, namely Logistic Regression (LR), Linear Discriminant Analysis (LDA), and Gaussian Naive Bayes (GNB) .We used scikit-learn libraries to import the three classification methods. The classification performance by these three methods were evaluated by 5-fold cross-validation. The experiments are conducted using Python 3.6.3 with 16-cores CPU.

We evaluate the classification models by using measure of Accuracy (ACC.), Recall (Rec.), Macro-averaged Precision (Prec.), and Macro-averaged F1-Measure (F1).

6.5 Classification Evaluation Results

We firstly evaluate how well the existing models can detect depressed tweets. After extracting the feature representations of each tweet, three different classification models are trained with the labelled data. We adopted a ratio of 75:25 to split the labelled data into training and testing data. We performed experiments by using TF-IDF, multi-modality, and combined features under three different classification models.

We compare the performance of models using only Multi-Modalities (MM) features with the three different classifiers as shown in Table 6.3. We found that Gaussian Naive Bayes for MM features obtained the highest Precision score compared to other classifiers, and Logistic Regression performs better than the other two classifiers in terms of Recall, F1-Score, and Accuracy.

Furthermore, in order to see how different features impact the classification performance, we use the TF-IDF standalone with three classifiers for comparison and the results are as shown in Table 6.4. It shows that all three classifiers can achieve satisfac-

Table 6.4: The performance of tweet depression detection based on TF-IDF only.

Features	Method	Precision	Recall	F1-Score	Accuracy
TF-IDF	LR	0.908	0.896	0.900	0.901
	LDA	0.906	0.893	0.897	0.898
	GNB	0.891	0.873	0.877	0.879

Table 6.5: The performance of tweet depression detection based on Multi-Modalties + TF-IDF.

Features	Method	Precision	Recall	F1-Score	Accuracy
MM+TF-IDF	LR	0.908	0.899	0.902	0.903
	LDA	0.912	0.899	0.903	0.904
	GNB	0.891	0.874	0.878	0.879

tory classification results. LR and LDA shared the highest Precision and F1-Score. LR competes with the other two classifiers according to the Recall and Accuracy.

Table 6.5 shows the model performance when we concatenate both MM and TF-IDF features, and we can see that the model has improved the performance further slightly. One conclusion we can draw here is that TF-IDF textual feature can make main satisfactory contribution to detect depression tweets, while other modality can provide additional support. This could be attributed to the lexical importance to depressed-related tweets. Another reason why the overall combination of multi-modal features did not give us a big lift in the results could be because, as mentioned earlier, tweets are highly sparse, poorly phrased, short texts, and noisy content. Therefore, deriving semantic content from tweets, for instance, using the LDA model would always be very challenging to get a huge boost in the results. This is mainly because any statistical model relies on the co-occurrence statistics which might be poor in our case. However, we still see an improvement in the overall recall result, which is important in this case, because we are noticing a reduction in the number of false negative detection. This is mainly because of the interplay between all three features which suggest that these features are important and cannot be ignored.

6.6 Detecting Depression due to COVID-19

After having testified our classification model, we utilise our approach to detect depressed tweets from different LGAs of NSW, Australia. Since our model deals only with English tweets, we had to exclude tweets in all other languages and input English tweets only

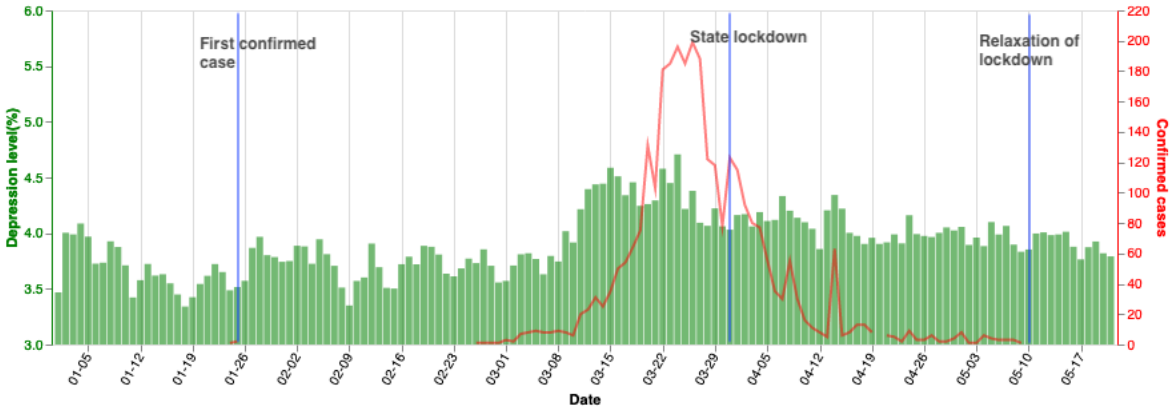


Figure 6.4: The community depression dynamics in NSW between 1 January 2020 and 22 May 2020.

into our model to predict depression. We ended up with 49 million tweets from 128 LGAs in NSW. We fed the LR with (MM + TF-IDF) model with these tweets, and the model found that nearly 2 million tweets were classified as depressed tweets. In this section, we show the depression dynamics in NSW during the study period between 1 January 2020 and 22 May 2020. The depression dynamics in different LGAs in NSW are also analysed to demonstrate how COVID-19 pandemic may affect people's mental health.

6.6.1 Depression dynamics in NSW

Fig. 6.4 presents the overall community depression dynamics in NSW with the confirmed cases of COVID-19 together during the study period between 1 January 2020 and 22 May 2020. "Depression level" refers to the proportion of the number of depressed tweets over the whole number of tweets each day. From this figure, we can find that people showed a low level of depression during the period before the significant increase in the confirmed cases of COVID-19 until 8 March 2020 in NSW. People's depression level was significantly increased with a significant increase in confirmed cases of COVID-19. The depression level reached to the peak during the peak outbreak period of COVID-19 on 26 March 2020. After that, people's depression level decreased significantly for a short period and then kept relatively stable with some short fluctuations. Overall, the analysis clearly shows that people became more depressed after the outbreak of COVID-19 on 8 March 2020 in NSW.

When we drill down into details of Fig. 6.4, it was found that people's depression was much sensitive to sharp changes of confirmed cases of COVID-19 on that day or after. For example, people's depression level had sharp changes around days of 10 March, 25

March, 30 March, and 14 April 2020 (there were sharp changes of confirmed cases of COVID-19 on these days). The sharp increase of confirmed cases of COVID-19 usually resulted in the sharp increase of depression levels (i.e. people became more depressed because of the sharp increase of confirmed cases of COVID-19).

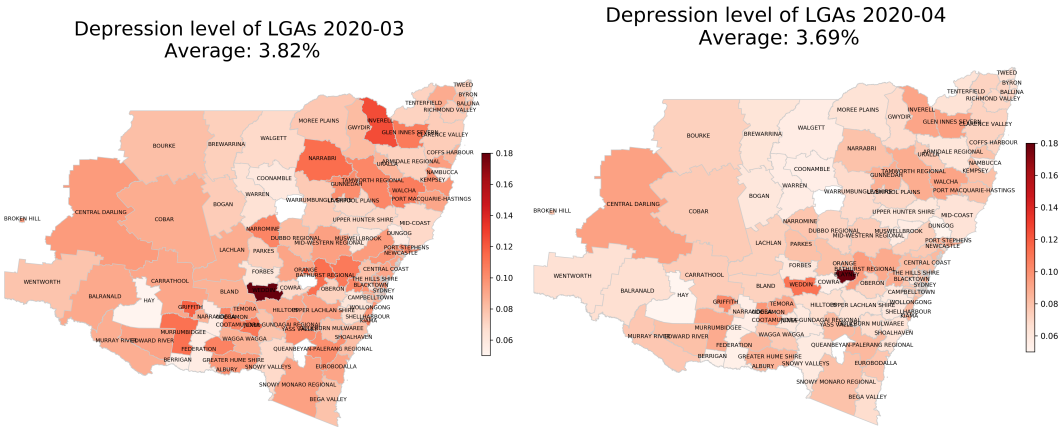


Figure 6.5: The choropleth maps of community depression in LGAs in NSW in March 2020 (left) and April 2020 (right).

6.6.2 Depression under implemented government measures and big events

This subsection investigates the links between people's depression and implemented government measures for COVID-19 (such as lockdown) as well as big events during the pandemic.

By investigating the labelled topics with the hashtag in the collected Twitter data, it was found that the topics of "lockdown" and "social-distancing" were started to be discussed actively from 9 March when the government encouraged people to increase social-distancing in the life, and the NSW government officially announced the state lockdown on 30 March and the restrictions were begun from 31 March[9]. The NSW government announced the ease of restrictions on 10 May[10]. When we link these dates with depression levels as shown in Fig. 6.4, it was found that people felt significantly more depressed when they started to actively discuss the lockdown restriction on 9 March. People were slightly more depressed after the official implementation of the state lockdown in NSW. The results revealed that the lockdown measure may make people

[9] https://gazette.legislation.nsw.gov.au/so/download.w3p?id=Gazette_2020_2020-65.pdf

[10] https://www.nsw.gov.au/media-releases/nsw-to-ease-restrictions-week-0

more depressed. However, people still became more depressed even if after the relaxation of lockdown. This is maybe because people still worried about the spread of this severe virus due to the increased community activities.

We are also interested in whether people's depression was affected by big events during the COVID-19 period. For example, the Ruby Princess Cruise docked in Sydney Harbour on 18 March 2020. About 2700 passengers were allowed to disembark on 19 March without isolation or other measures although some passengers had COVID-19 symptoms at that time, which was considered to create a coronavirus hotbed in Australia. The Ruby Princess Cruise has been reported to be linked to at least 662 confirmed cases and 22 deaths of COVID-19[11].

We link the depression dynamics as shown in Fig. 6.4 with the important dates of the Ruby Princess Cruise (e.g. docking date, disembarking date) and actual timeline of the public reporting of confirmed cases and deaths as well as other events (e.g. police in NSW announced a criminal investigation into the Ruby Princess Cruise debacle on 5 April) related to the Ruby Princess [11]. We have not found significant changes in people's depression on those dates. This may imply that the big events did not cause people's depression changes significantly.

6.6.3 Depression dynamics in LGAs

We further analysed community depression dynamics in LGAs in NSW. Fig. 6.5 shows examples of choropleth maps of community depression dynamics in LGAs in NSW for two different months which are March 2020 and April 2020. We can observe from the maps that the community depression level was different across different LGAs in each month. Furthermore, the community depression of each LGA changed in different months. On average, people in LGAs were more depressed in March than in April. This is maybe because the number of daily confirmed cases of COVID-19 was significantly increased to a peak and it was gradually decreased in April.

We also dig into more details of depression changes of LGAs around Sydney City areas. For example, Ryde, North Sydney, and Willoughby are three neighbouring LGAs in Northern Sydney. Their community depression dynamics and corresponding confirmed cases of COVID-19 are shown in Fig. 6.6, respectively. When comparing the dynamics in this figure, we can see that different LGAs showed different depression dynamics maybe because some events specifically related to that LGA. For example, in an aged

[11] https://www.theguardian.com/world/2020/may/02/australias-coronavirus-lockdown-the-first-50-days

care centre in Ryde LGA, a nurse and an 82-year-old elderly resident were first tested positive for coronavirus at the beginning of March[12]. After that, a number of elderly residents in this aged care centre were tested positive for COVID-19 or even died. At the same time, a childcare centre and a hospital in this LGA have been reported positive COVID-19 cases in March. Many staff from the childcare centre and the hospital were asked to test the virus and conduct home isolation for 14 days. All these may result in the significant community depression changes in March 2020 as shown in Fig. 6.6. For example, the depression level in Ryde LGA was changed significantly to a very high level on 10 March 2020 and 16 March 2020 respectively.

However, it was not found that the community depression dynamics in an LGA showed close relations with the dynamics of confirmed cases of COVID-19 in that LGA as we found in the state level. Maybe this is because the community depression dynamics in an LGA was affected largely by the confirmed cases in the overall state but not the local government area. This aligns with our common sense during the COVID-19 pandemic: even if our family is currently safe from COVID-19, we are still worrying about the life because of the continuing significant increases of COVID-19 all over the world especially in big countries.

6.6.4 Discussion

COVID-19 pandemic has affected people's lives all over the world. Due to social-distancing measures and other restrictions implemented by the government, people often use social media such as Twitter for socialising.

The results of this study showed that our novel depression classification model can successfully detect community depression dynamics in NSW state level during the COVID-19 pandemic. It was found that people became more depressed after the outbreak of COVID-19 in NSW. People's depression level was much sensitive to sharp changes in confirmed cases of COVID-19, and the sharp increase of confirmed cases made people more depressed. When we conducted a fine-grained analysis of depression dynamics in LGAs in NSW, our novel model can also detect the differences of people's depression in LGAs as well as depression changes in each LGA at a different period.

The study found that the policies/measures implemented by the government such as the state lockdown has had obvious impact on community depression within a short time period. The implementation of the state lockdown made people more depressed; however,

[12] https://www.smh.com.au/national/woman-catches-coronavirus-in-australia-40-sydney-hospital-st
html

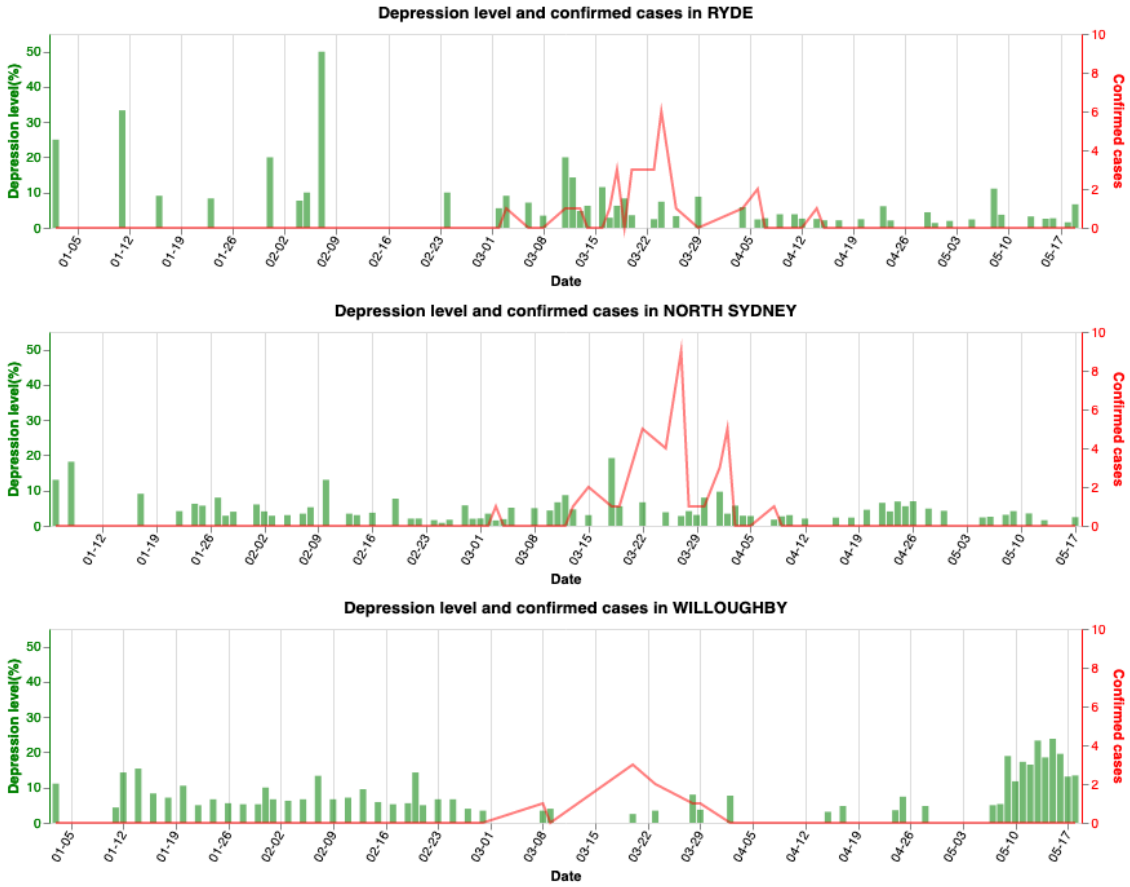


Figure 6.6: The community depression dynamics in Ryde, North Sydney, and Willoughby in Northern Sydney between 1 January 2020 and 22 May 2020.

the relaxation of restriction also made people more depressed. This could be primarily because people are still worried about the spread of COVID-19 due to the increased community activities after the relaxation of restrictions.

This study did not find the significant effects of big events such as the Ruby Princess Cruise ship coronavirus disaster in Sydney on the community depression. This is maybe because the data related to the disaster is small and passengers were also from other Australian states and even overseas besides NSW.

Since the proposed approach does not have constraints on locations of Twitter data, the proposed approach can be used to analyse the depression dynamics in other regions or countries. The proposed approach can also be generalised without special considerations to analyse depression dynamics over time and locations because of other events besides COVID-19.

6.7 Summary

This work conducted a comprehensive examination of the community depression dynamics in the state of NSW in Australia due to the COVID-19 pandemic. A novel depression classification model based on multi-modal features and TF-IDF was proposed to detect depression polarities from the Twitter text. By using Twitter data collected from each LGAs of NSW from 1 January 2020 until 22 May 2020 as input to our novel model, this chapter investigated the fine-grained analysis of community depression dynamics in NSW. The results showed that people became more depressed after the outbreak of the COVID-19 pandemic. People's depression was also affected by the sharp changes in confirmed cases of COVID-19. Our model successfully detected depression dynamics because of the implementations of measures by the government. When we drilled down into LGAs, it was found that different LGAs showed different depression polarities during the timeframe of the tweets used in our study, and each LGA may have different depression polarity on different days. It was observed that the big health emergencies in an LGA had a significant impact on people's depression. However, we did not find significant effects of the confirmed cases of COVID-19 in an LGA on people's depressions in that LGA as we observed in the state level. These findings could help authorities such as governmental departments to manage community mental health disorders more objectively. The proposed approach can also help government authorities to learn the effectiveness of policies implemented.

In this special period of the COVID-19 pandemic, we focused on the effects of COVID-19 on people's depression dynamics. However, other factors such as unemployment, poverty, family relationship, personal health, and various others may also lead people to be depressed. Our future work will investigate how these factors may affect community depression dynamics. Furthermore, community depression will also be investigated using the topics over time model and using the temporal topics as multi-modal features. More recent and advanced classification models will be investigated to classify people's depression polarities.

CHAPTER 7

DEPRESSION DETECTION AT USER-LEVEL AND ITS IMPACT DURING PANDEMIC

7.1 Background and Motivation

The novel coronavirus has quickly spread throughout the globe, affecting 218 nations and territories as depicted in Figure 7.2. By 31 July 2021, there were more than 4.2 million deaths and more than 197 million confirmed cases of illness.[1]. This new pandemic has created severe health and socio-economic disruptions worldwide. The World Health Organization (WHO) claims that COVID-19 is a pandemic that is taking lives throughout the globe every day, and the severity of the outbreak rates strongly supports this claim. The World Health Organization reported a new coronavirus outbreak in Wuhan, China, on 31 December 2019 (WHO) (160). COVID-19 is spread by droplets and fomites during close, unprotected contact between sick and affected people who are not properly treated[2].

This epidemic has influenced mental health and the physical dangers posed by the virus (203). The pandemic has not only caused complications by physically affecting the individuals but also impacted their mental lives. The lockdown and stay-at-home policies have also affected different types of individuals; for instance, a recent study found that

[1]https://covid19.who.int/

[2]https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf

Figure 7.1: A sample of depressed user tweets during the first months of COVID-19.

lockdown could lead to more violence in the UK[3], which will have a detrimental societal impact. After staying at home for very long, quickly lifting lockdown restrictions will have an impact on the mental health too[4]. According to some studies (57; 74; 159; 27), the world's population will face many psychological problems related to anger, anxiety, and others due to their isolation in homes. These studies point out that people can only get out of this psychological state initially through the intervention of governments in psychological support during crises, not only treating them after disasters are over. With the spread of COVID-19, mental health problems are suspected of increasing among people. Other cascading factors that could severely impact people's mental health are job losses, increasing prices, non-availability of certain food items, and others.

Mental health distress, including peri-quarantine anxiety, depression, and phobias specific to COVID-19, is supposed to increase during the pandemic, affecting individuals and calling for intervention. Due to pandemic-induced social distancing protocols, health interventions have become the most common form of psychosocial support for mental health. However, there is no evidence of the effectiveness of remote psychosocial interventions in improving mental health outcomes. Infection control strategies such as lockdown and social distancing (physical distancing) try to slow disease spread by decreasing direct contact among people. Conversely, lockdown, quarantine, and social isolation may lead to personal freedom loss, future uncertainty and anxiety, as well

[3]https://www.msn.com/en-gb/health/mindandbody/youth-violence-likely-to-explode-over-summer-uk-experts-fear/ar-AAMtHuO?ocid=msedgntp

[4]https://www.msn.com/en-gb/health/mindandbody/lifting-restrictions-quickly-is-just-as-detrimental-to-our-mental-health-as-lockdown/ar-AAMvF5f?ocid=msedgntp

as a large increase in the incidence of emotional diseases and mental health problems (154; 200). Lockdown measures are connected to higher than average incidence of mental illness such as depression, according to a study by Rossi et al., (150). More extended quarantine periods, fears of illness, frustration, boredom, insufficient supplies, lack of knowledge, financial loss, and stigma were all stressors. As a result, Tull et al., (180) found out that remaining at home was linked to higher levels of health anxiety, financial anxiety, and loneliness among healthy people.

In the midst of the COVID-19 epidemic, social media usage has increased as more people rely on acquiring the newest concerning COVID-19 information (136). Additionally, social media, including a variety of internet services and platforms like Facebook and Twitter, allow users to converse, interact and different information sharing. While social media can help distribute information, which could be beneficial in the fight against the epidemic, it has also been related to anxiety and depression (118). We hypothesize that the COVID-19 pandemic and its social restrictions may affect depressed users and thus may be reflected in their daily tweets. Figure 7.1 shows a sample of tweets from a depressed user during this pandemic. From the Figure, we notice that the user continues to post about their mental health issue, and we believe there is an increase in the proportion of depression-related tweets during the pandemic compared to before. In April 2020, several countries performed various lockdown restrictions and testing strategies to contain the virus's spread (83), and these restrictions were reflected in user tweets, as we observe in Figure 7.1. One of the recent studies by Hua et al., (68), shows that many people use social media to disseminate terrifying COVID-19-related information, such as information on individuals who passed away as a result of COVID-19 or who are battling the disease; hence, many of these posts could be shared via depressed people.

In January 2020, lockdowns for the COVID-19 pandemic began in China (83). They were quickly followed by lockdowns in many countries worldwide, including Egypt and Germany. The first lockdown, crucially, was generally unanticipated by the general public in all countries. As a result, not everyone can tolerate the limits in daily life and behavior equally well. WHO and health care professionals have recommended that counseling programs supporting and directing individuals in behavior modification become part of the COVID-19 pandemic preventive measures to avoid extra mental health costs in the general population (126). However, to properly promote mental health, well-being, and behavior, a better scientific knowledge of how individuals perceive and mentally react to the present COVID-19 epidemic is needed. A better scientific understanding of the social aspects of life most affected by the current COVID-19 pandemic, including

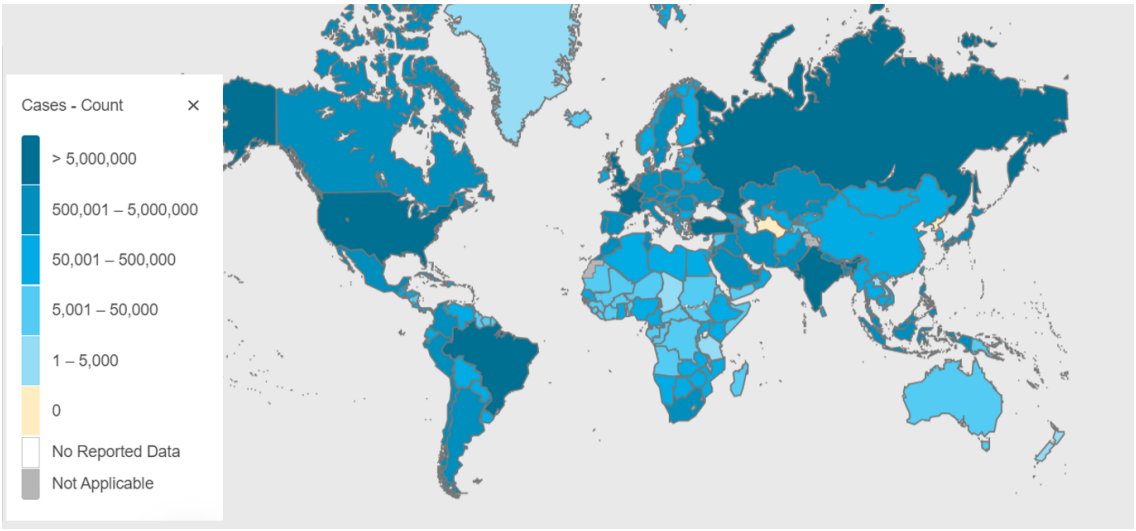


Figure 7.2: As of July 23, 2021, the following countries or geographic locations have confirmed cases of COVID-19.

how individuals think, feel, suffer, deal with situations, and perceive a threat, emotion, and emotion behavior control (183). Moreover, a depressed user may post in various linguistic styles, using depressing words, antidepressant mentions, and descriptions of depressed symptoms, all of which can help diagnose depression. While tweets may contain rich information sources, and due to their unstructured nature, obtaining insights from them can be difficult and time-consuming. Organizing text data is becoming ideal due to advancements in natural language processing (NLP) (103). In particular, text classification is among the most fundamental techniques in machine learning. In text classification, given annotated training data, the model learns patterns from this data. Given some unseen data, the goal of the model is to assign labels to each instance; annotations or labels could be in the form of topic, genre, or sentiment (10; 128; 181). The importance of text classification has never been greater than it is now, especially for companies seeking to boost productivity or profitability. Most recently, new deep learning models have achieved SOTA performance in text classification. More recently, transfer learning methods have become very popular. In this setup, a model is first trained on some general text collection. The model is then fine-tuned on task-specific problems in certain domains. Several models already have been developed that belong to such as class of transfer learning model, for instance, ULMFit fine-tunes a 3-layered LSTM (66) and Bidirectional Encoder Representations from Transformers (BERT) from Google (41) has been shown to be effective in several downstream tasks. Nevertheless, classifying tweets in social media is still challenging due to the structure of the user tweets and the

fact that these text instances are very short and sometimes ambiguous.

Recently, the hierarchical attention networks (HAN) model proposed by Yang et al., (210) achieved SOTA results on text classification. What differentiates HAN from other existing approaches to document classification is that it exploits the hierarchical nature of text data by dividing a document into two levels, word-level hierarchy and sentence-level hierarchy. The authors proposed a new model that maintains a hierarchical structure consisting of word and sentence encoders. The word encoder extract features at the word level and sends them to the sentence encoder for processing. The sentence encoder extracts the features at the sentence level and predicts output probabilities at the final layer. However, HAN takes substantially longer to train than CNN-based techniques because they utilize RNNs. Long-distance linguistic elements are more difficult to capture using CNNs. Despite this shortcoming, CNNs are typically just as efficient as RNNs in performing many basic NLP tasks. The reason why we developed our novel model based on HAN is that user posts may include linguistic cues at various levels of word and tweet levels. To identify a depressed user on social media, every word in a tweet and every user's tweet is equally essential Figure 7.1.

We propose a new hierarchical convolutional neural network (HCN) model to better model user tweet classification through social media. Our model can extract meaningful feature representations from the word and tweet levels. Eventually, we studied our model performance using a multi-channel convolution neural network CNN and multilayer perceptron MLP in word-level encoding to combine the advantages of two traditional neural network models. We found our model performance increased when we used two channels for word encoding, and we call this model with two channels HCN+. Our research makes the following keys contributions:

- We develop a depressed user classification model based on a hierarchical convolution neural network **(HCN)** and hierarchical attention that can learn and extract linguistic relationships and local features from tweets.

- Besides HCN, we present a variant of hierarchical convolution neural network **HCN+**. The model (HCN+) expands word encoding by using two channels and reading the input differently in parallel using multilayer perceptron and convolutional neural network to extract various features from the same input and boost our model performance to identify depressed Twitter users.

- We collected a dataset, including posts from users with and without depression during COVID-19, to identify and analyze online depression during the pandemic.

We are aware that this is the first study evaluating the behavior of depressed users during the time of the COVID-19 epidemic. Furthermore, we have made this dataset publicly available to aid in the research on mental health and well-being[5].

7.2 Preliminary on Depression Detection at User-Level and Its Impact During Pandemic

In this section, we summarise some closely related work, including our previous works. We mention how the model developed in this work is substantially novel and different from our previous works.

7.2.1 Depression Detection on Social Media

The majority of prior research, including ours, have focused on user behaviour to identify whether the user is depressed or suffering from another mental disease. Deep learning approaches for online depression detection, such as public discussion forums, have been developed to model data from the social media. Most methods take words and word n-grams as user-level features and apply classic classification algorithms such as Support Vector Machines (SVM).

Users that were depressed had much greater ratios of negative emotion phrases, according to Tsugawa et al (178). In general, the linguistic style of postings may be used to extract features (67; 213; 213). Many factors, such as user tweets, replies, retweets, post time, emotions, and so on, might help identify depression. Shen et al. (156) used Twitter to create a well-labelled depression and non-depression dataset and then investigated a variety of depression-related features and developed a multimodal depressive model to identify depressed users. They grouped these features into six depression-related features, including clinical depression criteria and online social media behaviours.

Wang et al., (194) proposed a way for automatically gathering people who described themselves as having an eating disorder in their Twitter profile descriptions to detect eating disorders within social media communities. The authors collected features of linguistic from users for psychometric qualities, and they used similar settings described in (139; 85; 157). From Twitter and Weibo, the authors collected 70 features. They took these characteristics from a user's profile, and user engagement characteristics like

[5]https://github.com/hzogan/Post-COVID_dataset

many followees and followers. Wong et al., (194), on the other hand, integrated user-level and post-level semantics and framed their problem as many instances learning setup; this approach has the benefit of learning from user-level labels to post-level labels identification (202). Moreover, due to the need to describe the underlying workings of a deep learning system and make them more dependable and understandable, recently explainable machine learning has drawn a lot of attention (210; 29; 23). Inspired by that, Zogan et al., (226) recently introduced a model that improves depression detection by incorporating the explainability of depressed user tweets. Their model interplay between multilayer perceptron (MLP) and hierarchical attention network (HAN). MLP was used to encode users' online behaviour, while HAN encoded all user tweets at two levels: word-level and tweet-level. They determined each tweet and word weight and extracted characteristics derived from user tweets' semantic sequences. Another work by (224), the authors argued that using all user tweets to identify a depressed user is ineffective and could even degrade a model performance; therefore, they proposed a new summarization framework interplay between extractive and abstractive summarization to generate a shorter representation of user historical tweets and help to reduce the influence of content that may not eventually benefit the classifier.

7.2.2 Depression Detection due to COVID-19

COVID-19 has had a negative influence on people's mental health, and as a result of the epidemic, many people are experiencing an increase in mental health issues. Various studies have recently reported this. For instance, Zhou et al., (223) developed a depression classification model that extracts multimodal features from topic, emotion and domain-specific viewpoints using the tf-idf algorithm. Their study focuses on detecting community level depression in Australia and local areas only during the pandemic. Their findings reveal an increase in depression following the COVID-19 epidemic. While in our new study, we will analyse tweets of users with and without depression during eight months before and after the start of the COVID-19 pandemic. Galea et al., (52) claimed that the pandemic's mental health repercussions are applicable in the short-term and long-term, and that COVID-19 provided an opportunity to explore overseas students' behavioural responses to the imposed limits on travel and social interaction. Due to social distancing, the current COVID-19 epidemic substantially impacts daily activities, including work, social, educational, and recreational activities. Some of these situations may raise the risk of aggressiveness and suicide (69; 218). During the pandemic period in China, Huang et al., (69) investigated anxiety disorder, depression, and sleep quality and

found that the participants had varying degrees of mental health difficulties. According to the same study results, 35% of participants had anxiety, 20% had depression, and 18% had poor sleep quality.

Our proposed approach, on the other hand, focuses on automatically identifying depression online, which has the potential to monitor people's mental health. We have also developed a novel model for automatic modelling on user's depression as a result of the pandemic.

7.3 Dataset

To study depressed users during the COVID-19 pandemic, we have created a new dataset that includs users tweets pre- and during COVID-19. We show in Table 7.1 a summary of this dataset.

	Pre-COVID-19		**COVID-19**	
	# User	# Tweet	# User	# Tweet
Depressed	2,159	447,856	741	326,129
Non-Depressed	2,049	1,349447	682	931,527

Table 7.1: Summary of the datasets that we used in our research

7.3.1 Pre-Covid-19 Dataset

Shen et al., (156) constructed a textual depression dataset on Twitter. The authors labeled users as depressed if they found a user's tweet that contained a specific pattern. Between 2009 and 2016, they constructed depressed users based on the content of their posts. There are around 2K depressed users and around 400K tweets in all. The tweets for non-depressed were collected in December 2016, including over 2K users and over a million tweets.

7.3.2 COVID-19 Dataset

Here, we provide a description of our dataset, which was generated following the COVID-19 pandemic started to penetrate across different nations. according to the tweets' IDs in (156), first, we constructed user tweets after COVID-19 using Twitter APIs. We set up a time-frame to study all users in the COVID dataset from the 1st of September 2019 until the 20th of April 2020. We crawled tweets of 1423 users who were tweeting during the

time-frame of our study. Finally, we constructed this novel dataset with over a million and two hundred tweets from both depressed and non-depressed users based on these tweets.

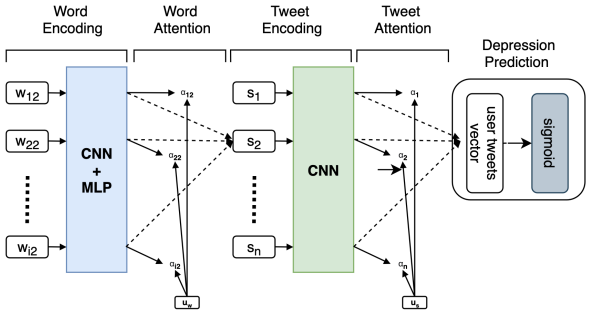


Figure 7.3: A diagram of (HCN) that we employed for user all user posts

7.4 Our Proposed Model

Figure 7.3 depicts the hierarchical attention network that was provided to learn user tweets representation of as we were inspired by (210). As mentioned above, there are some key differences between our model and the HAN model. One difference is that we have used the CNN model in our framework with two channels for word encoding. Assume that U is a user who posted M tweets $T = [t_1, t_2, ..., t_M]$ each of which had N_i words $t_i = [w_1, w_2, ..., w_N]$. The series of d-dimensional embeddings of each tweet's words, $w_i = [w_{11}, ..., w_{MN}]$, serves as its representation. Each word is represented by a fixed-size vector from pre-trained word embeddings as the input layer. Figure 7.3 depicts the overall structure of our HCN+.

7.4.1 Word Encoding

Given a tweet with words w_{it}, $t \in [0, T]$, we begin by converting the words to vectors using an embedding matrix. Each word vector is represented by a fixed-size vector obtained using pre-trained word embeddings as the input layer. To initially capture the contextual information of the annotations, CNN is employed as the word level encoder. We utilized one-dimensional convolution, where a filter vector slides across a sequence and detects features at different points in one-dimensional convolution. The convolution kernel completes the convolution operation in the k-dimensional window, the convolution kernel $w \in \mathbb{R}^{hk}$ operates on the input data matrix $x_{i:i+h-1}$ to create the feature c_i each time, and the output features are as follows:

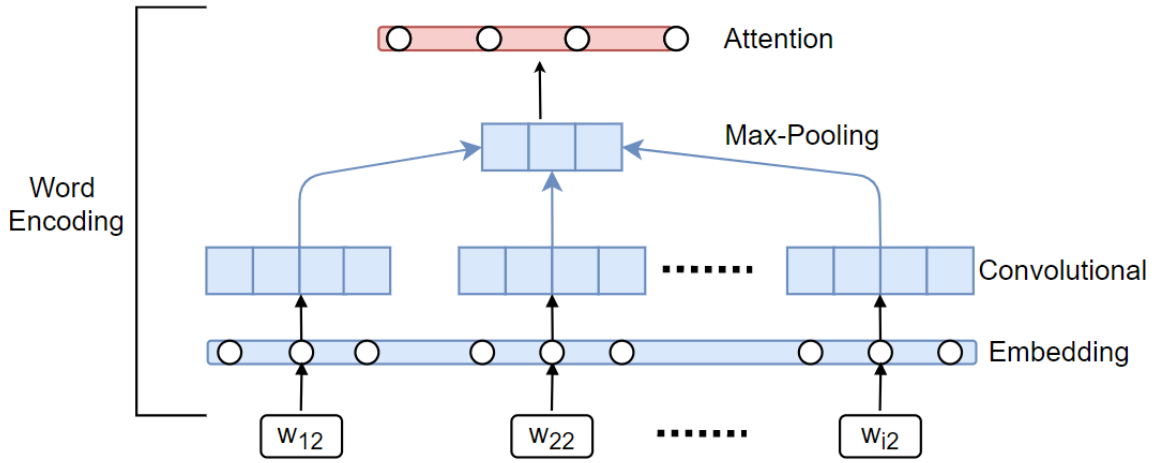


Figure 7.4: An illustration of one channel CNN model that we use for HCN word encoding

(2) $$c_i = f(wx_{i:i+h-1} + b)$$

where f is the nonlinear activation function $relu$, h is the sliding window range, b represent the bias and $x_{i:i+h-1}$ is the result of joining the word x_i to x_{i+h-1}. As a result, the convolution kernel is applied to the tweet sequence of length n, yielding $n-h+1$ results:

(10) $$c = \{c_1, c_2, \cdots, c_{n-h+1}\}$$

Reducing the size of features is done using the pooling layer and eliminate redundant information. The pooling process is carried out using the pooling technique. The K-max pooling layer is utilised in our architecture to down-sample the local feature maps C and extract global feature representations of short texts with fixed-length, as follows:

(3) $$\hat{c} = \max(c)$$

The attention mechanism is then described. It is important to create a trainable and expected to capture global word vector u_{ij} for all words. The $\hat{c}_{ij}$ annotations build the basis for attention, which begins with another hidden layer. The annotations u_{ij} will be represented as following:

(7.1) $$u_{ij} = \tanh(W_w \hat{c}_{ij} + b_w)$$

Where W_w and b_w are the trainable parameters that learned by the model after random initialisation. Then the product $u_{ij}u_w$ (u_w is randomly initialised) is intended to emphasise the significance of the word j and normalised by a softmax function to an importance weight per word α_{ij}:

$$\alpha_{ij} = \frac{\exp(u_{ij}u_w)}{\sum_j \exp(u_{ij}u_w)} \tag{7.2}$$

A weighted concatenation of word representations and the previously determined annotations is known as the tweet vector v_i, where α_t is the importance weight per word:

$$v_i = \sum_t \alpha_{ij}\hat{c}_{ij} \tag{7.3}$$

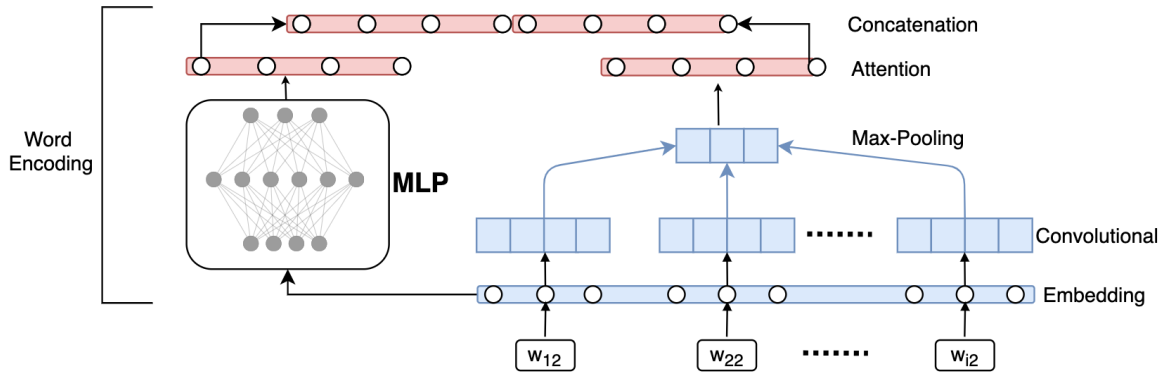


Figure 7.5: An illustration of two channel CNN+MLP model that use for HCN+ word encoding

For our first model HCN, we have utilized one channel CNN for word encoding Figure 7.4 depicts the word encoding component for HCN. Then, in order to combine the advantages of two traditional neural network models and alleviate their shortcomings, we have utilized two channels CNN and a multilayer perceptron (MLP), to obtain multi-channel representations. The multi-channel representations reflect the various contributions of different words to a tweet's semantics and provide a way to represent a tweet from several views. Hence, we have examined the effectiveness of using multi-channel CNN-MLP for the word encoding of our second model HCN+ depicted in Figure 7.5. Each deep learning model has its own method for converting target data into feature vectors. The embedding layer first fed into the MLP with the attention model, which generates an MLP feature vector and fed CNN with the attention model, generating a

CNN feature vector. These vectors are concatenated and produced a tweet embedding representing the content within that tweet.

7.4.2 Tweet Encoder

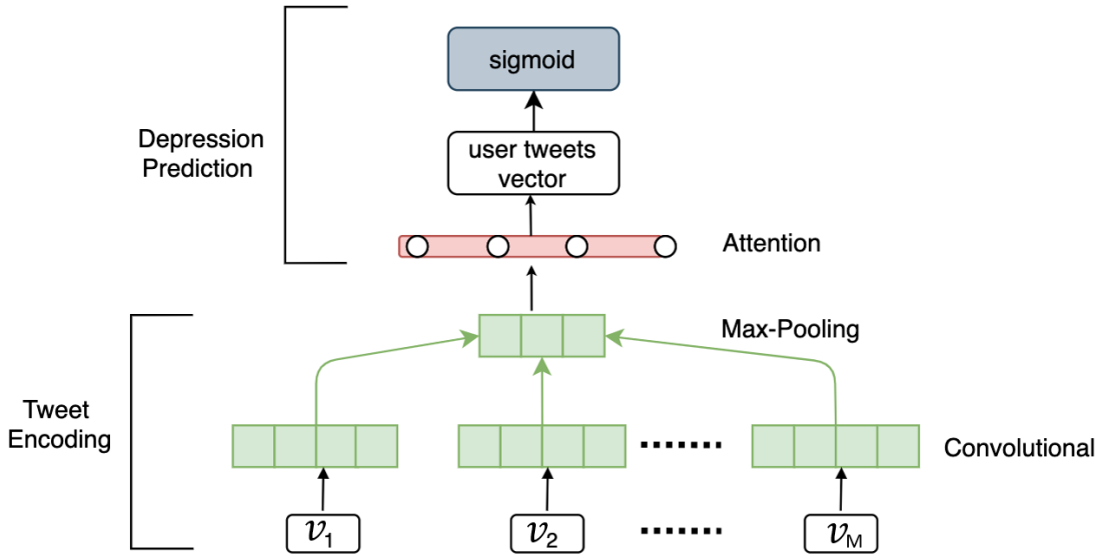


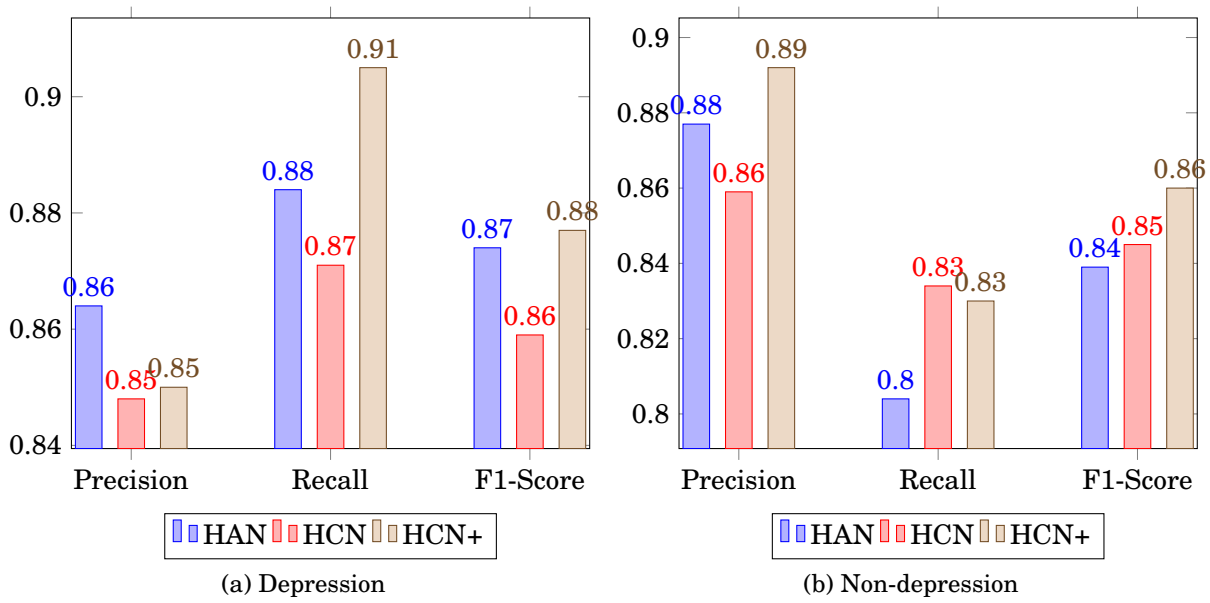
Figure 7.6: An illustration of a tweet encoder network

Yang et al. (210), achieved SOTA performance on HANs by employing a hierarchical framework that breaks down texts into sentences first. The word-level reads in word embeddings from a given sentence. It produces a sentence embedding representing the content within that sentence. In comparison, the tweet-level reads in the tweet embeddings produced by the word-level as depicted in Figure 7.6. Given that a user's tweets might not be significant in identifying and elucidating a depression of a person, the attention layer is employed with the tweets of user to obtain the more important tweets. Utilizing the tweet level attention layer, we will gather the connected tweets. The importance of the i tweet is predicted to be indicated by the product $u_i u_s$, which has been standardised to a α_i importance weight for each tweet. Eventually, a vector called $\hat{t}$ will be used to compile all the tweet data from user posts:

$$\hat{t} = \sum_t \alpha_i h_i^t \tag{7.4}$$

Table 7.2: Performance Comparison on Pre-COVID datasets. HCN+ outperforms baselines.

Training Data	Model	Precision	Recall	F1-Score	Accuracy
Tweets Summarization	XLNet (base)	0.889	0.808	0.847	0.847
	BERT (base)	0.903	0.77	0.831	0.837
	RoBERTa (base)	**0.941**	0.731	0.823	0.836
	BiGRU-Att	0.861	0.843	0.835	0.837
	CNN-Att	0.836	0.829	0.824	0.824
	CNN_BiGRU-Att	0.868	0.843	0.848	0.835
All user tweets	BiGRU	0.766	0.762	0.786	0.764
	CNN	0.817	0.804	0.786	0.806
	HAN	0.870	0.844	0.856	0.835
	HCN	0.853	0.852	0.852	0.852
	HCN+	0.871	**0.868**	**0.869**	**0.869**



(a) Depression

(b) Non-depression

Figure 7.7: Comparison between HCN+ and other hierarchical text classification models (a) for depression prediction and (b) for nod-depression prediction

7.4.3 Classification Layer

At the classification layer, we must determine whether or not the user is depressed. So far,we've covered the encoding of user tweets (s) by simulating the tweet and word levels of the hierarchical structure. Such a network's output is often sent to a sigmoid layer for classification:

$$\hat{y} = \text{Sigmoid}(b_f + \hat{t}W_f) \tag{7.5}$$

where $\hat{y}$ denotes the predicted probability vector. Predicted probabilities of labels being 0 (non-depressed user) and 1 (depressed user), they are correspondingly represented by the symbols $\hat{y_0}$ and $\hat{y_1}$. Afterward, the ground-truth label for each user y, we seek to minimise the cross-entropy error:

$$\text{Loss} = -\sum_i y_i \cdot \log \hat{y}_i \tag{7.6}$$

y_i represents the user who has the ground truth label (either non-depression or depression) and $\hat{y}_i$ is the predicted probability.

7.5 Experiments and Results

The experimental results of our model are shown in this section and compares them with different comparative models. We also present qualitative results.

7.5.1 Experiment Setup

7.5.1.1 Dataset

To evaluate the effectiveness our models, we conduct our experiments on Shen et al., (156) pre-COVID dataset, as shown in Table 7.1, which contains users and their posts on Twitter. Each user is labelled either depressed or non-depressed. For preprocessing, we remove users who have less than ten tweets and for evaluation, we randomly split the dataset into training and test set with a ratio of 80:20 with 5-fold cross-validation.

7.5.1.2 Comparative methods

This section presents an experimental assessment to verify HCN's performance. We looked into three well-known pre-trained models since they are frequently used in contextual language models created using recent deep learning techniques. Our model is compared with different classification methods using the following data inputs:

1. **Tweets Summarization**: Zogan et al, (224) summarize all user tweets utilizing a new summarization framework interplay between extractive and abstractive summarization to generate a shorter representation of user historical tweets and help

to reduce the influence of content. Their experiments for summarization sequence classification have examined several models, Conventional neural network with attention CNN-Att, Bidirectional Gated Recurrent Neural Network with Attention BiGRU-Att. Three pre-trained models for transformers have also been investigated by the authors, and they are XLNet(209), BERT (41) and RoBERTa (96).

2. **All User Tweets**: For all user tweets, our model is comapred to the following classification methods:

    - **BiGRU**:For the purpose of classifying user tweets, we employed the **Bidirectional-GRU**(26) with attention method that we deployed to get user tweet representations.
    - **CNN**: In order to represent user tweets and capture the semantics of various convolutional window sizes for depression detection, we used **CNN**(77) with an attention mechanism.
    - **HAN**:In order to identify depression in user postings, a hierarchical attention neural network architecture(210) is deployed. The network encrypts first user postings by paying attention to both the words in each tweet and the tweets themselves. Bidirectional-GRU is used (GRU).
    - **HCN**: Similar to HAN, however instead of utilizing (GRU), hierarchical convolutional network (HCN) rely on architectures based off convolutional neural networks.
    - **HCN+**: The proposed model in this chapter.

7.5.1.3 Evaluation metrics

We employ the standard metrics for information retrieval such as accuracy, F1-score, and precision as metrics to examine the classification performance. These metrics are widely used in previous works for depression detection (156; 226; 224; 223).

7.5.1.4 Experimental settings

We performed the experiments with Python 3.6.3 and Tensorflow 2.1.0. Word embeddings is initialized by Glove (134). The dimension of word embedding is 100, and the dropout rate is set to 0.5. We created 100 different filters for the convolutions, each with length 4, so the result will bring 100 different convolutions. We used the Adam optimization

algorithm for both HCN and HCN+ with default value learning rate (lr) = 0.001. We train HCN+ for 20 epochs on all the data with a batch size of 32.

7.5.2 Results

In this section, we report the quantitative results obtained from different models. Evaluation results for different competing methods are presented, where the best results for the best model are highlighted in bold in Table 7.2. The first part of the table shows the effective results of using summarization sequences of user posts to detect depression; the performance compared using some models that achieved a new state-of-the-art result on many NLP tasks, such as text classification. We see that CNN_BiGRU-At outperforms other models with F1-score and recall; however, XLNet and RoBERTa perform best among all the different models with accuracy and precision, respectively.

The second part of Table 7.2, shows the results of using all user tweets; we observed that all the hierarchical text classification models (HAN, HCN and HCN+) efficient outperform other neural network-based methods, such as BiGRU and CNN. Furthermore, we observed that our two models, HCN and HCN+, outperform other models in terms of Accuracy and Recall. Comparing our model with HAN, Our model HCN+ boosts about 3.4%, 1.3%, 2.4% and 0.1% in terms of Accuracy, F1-score, Recall and Precision.

Generally, our models based on the hierarchical network can consistently outperform other methods in terms of Accuracy, F1 Score and Recall on both training data (tweets summarization and all user tweets). Our models based on hierarchical networks achieve a relative improvement of 0.5% for HCN and 2.2% for HCN+, compared against the best results (XLNet) in terms of Accuracy.

Finally, to better understand and examine the effectiveness of our model, we have compared the performance of our model (HCN+) with other hierarchical text classification models (HAN and HCN) in order to predict the depressed (Figure 7.7-a) and non-depressed-users (Figure 7.7-b). We found out HCN+ obtains better performance in both labels in terms of F1-score, showing 87% and 86% for depressed and non-depressed users, respectively.

7.5.3 Discussion

To study user tweets dynamic during COVID-19, we set up a time window to study all users in the pre-COVID dataset from 1 September 2019 until 20 April 2020. According to the table, we studied 1423 users; 741 were labeled as depressed users before COVID-19.

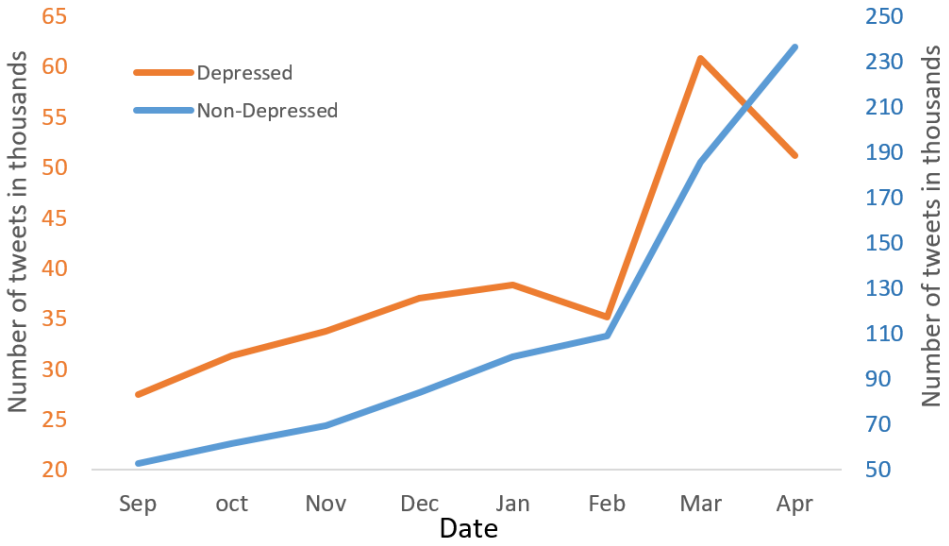


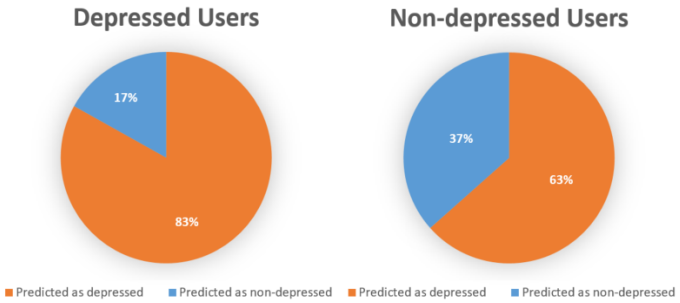
Figure 7.8: Monthly tweets for all users during the COVID-19

Figure 7.8 shows the monthly users' posts during our study window. We see that tweets for both depressed and non-depressed users have increased per month since the first cases of COVID-19 were reported (according to Lancet journal (160)). We have also observed that in February 2020, the number of tweets increased dramatically among all users.

After having testified our classification model, we utilize HCN+ to users in the COVID dataset (Table 7.3). First, we studied all user-collected tweets in the time frame from September to April and showed the proportion of positive cases among depressed and non-depressed users during COVID-19 Figure 7.9. As shown in Figure 7.9a, more than 80% of depressed users were predicted via HCN+ as depressed during the four-months time frame. The proportion in Figure 7.9 looks natural since these users have been diagnosed with depression before COVID-19. However, Figure 7.9b shows that 63% of non-depressed users were predicted as depressed. It shows how COVID-19 impacted normal users' tweets during the pandemic.

We further analyzed monthly dynamic user tweets. Figure 7.10 shows monthly dynamic depressed user tweets during our study time frame. We observed the number of depressed users predicted as positive, roughly the same during the first three months. However, once the first case was reported in late December 2019 in Wuhan (160), we can see that the number of positive cases decreased in December, which may be due to depressed users posting news tweets instead of tweets to expose their feelings. We also noticed that the number of positive cases of depressed users reached a peak in March,

which is the same month that World Health Organization WHO declared that COVID-19 is a pandemic (35). Also, in March, several countries performed various lockdown restrictions and testing strategies to contain the virus's spread (83).



(a) Depressed Users (b) Non-depressed Users

Figure 7.9: The proportion of positive cases among depressed and non-depressed users during COVID-19.

On the other hand, non-depressed dynamic tweets in the first four months of the study time-frame are quite similar to depressed users' tweets during the same time frame Figure 7.10. Like depressed users, the number of positive cases of non-depressed users also peaked in March. However, we noticed something considerably interesting, the portion pattern of depressed users' positive cases before and after December was slightly the same. While in non-depressed users, as we can see in Figure 7.11, the number of positive cases rate starting to increase in January compared to the positive cases rate in

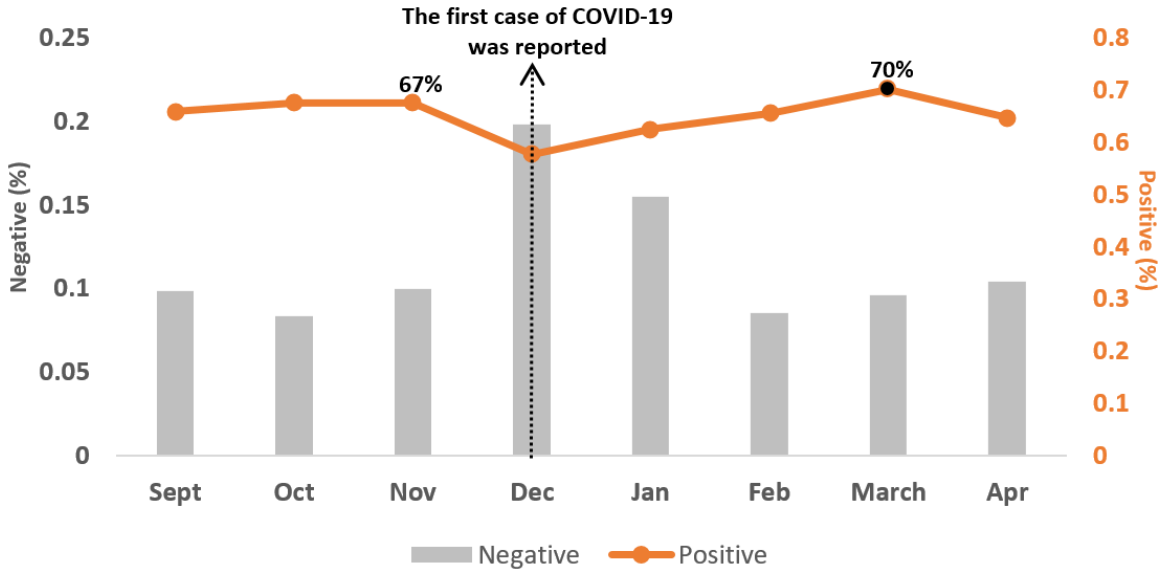


Figure 7.10: Depressed user dynamics between September 1, 2019, and April 20, 2020

November, which is the month before the first case of Corona was announced. We found out the positive cases rate increased by 3%, 9%, 15% and 12% in January, February, March and April, respectively.

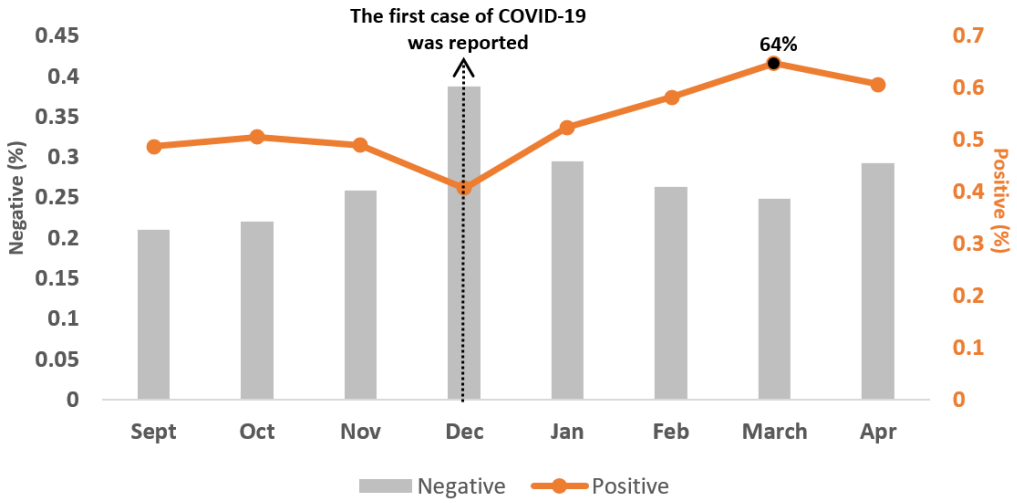


Figure 7.11: Non-depressed user dynamics between September 1, 2019, and April 20, 2020

7.6 Summary

In this chapter, we studied tweets of depressed and non-depressed users during eight months before and after the start of the COVID-19 pandemic. A user classification model to automatically detect depressed users based on a hierarchical convolution neural network (HCN) is proposed, which exploits data from Twitter. HCN considers the hierarchical structure of user tweets (tweets-words) and contains an attention mechanism that can find the most crucial tweets and words in a user document while also considering the context. We expand word encoding by using two channels and read the input in different ways in parallel using MLP and CNN to extract different features from the same input and boost our model performance. The results showed that our two models (HCN and HCN+) outperform strong comparative models and effectively detect depressed users. Furthermore, we have also investigated depressed users' well-being during COVID-19 by crawling and classifying their tweets during the pandemic. Moreover, we have analyzed depressed users' well-being during COVID-19 by crawling and classifying their tweets during the pandemic. We found that the COVID-19 pandemic and its restrictions, such as lockdowns and changes in the workplace, impacted many

depressed and non-depressed users. In the future, besides the users' tweets, we will analyze user behaviors related to depression during the COVID pandemic, such as social engagement and social interaction with others. This would provide the model with more contextual information and allow us to concentrate on a task where our model not only detects depression but also automatically gives a possible diagnosis. Moreover, we will aim to detect a user's loneliness during the pandemic, which is one mental illness that has never been before in this field. Loneliness is considered one of the early depression symptoms; therefore, its detection will help in early depression detection.

CHAPTER 8

CONCLUSION AND FUTURE WORK

8.1 Contributions

To solve the problem of depression detection in social networks, some solutions are proposed in this thesis. The conclusions and main contributions of the works in this thesis are listed:

- We have proposed a novel hierarchical deep learning network that coalesces multiple fully connected layers to integrate user behavioural representation and user posting because automatic summarization gives our computational framework a natural advantage that allows it to concentrate only on the most relevant info during model training and significantly reduces the curse-of-dimensionality problem (history-aware posting temporal network). In order to automatically choose the most relevant user-generated information, our framework for depression detection summarises the relevant user post history first.

- By extracting features from user behavior and the user's online timeline, we have developed explainable Multi-Aspect Depression Detection with Hierarchical Attention Network (MDHAN) for identifying depressed people through social media analysis (posts). We utilized a real-world data set with people who were both depressed and not depressed. Our key contribution is a brand-new hybrid computational model that not only accurately models real-world data but also

contributes to their interpretation. To determine the importance of each tweet and phrase and to extract semantic sequence characteristics from user timeline posts, we assign the multi-aspect attribute that reflects user behaviour into the MLP and user timeline posts into HAN (posts). According to our analysis, utilizing this hybrid network improves the accuracy of classifying users and particularly excels in identifying individuals with depression, outperforming other established methods. Additionally, it provides sufficient explaining for its predictions.

- we studied a novel problem of modeling narrative elements in social media to analyze how our posts can be used to understand our narrative. A user hybrid classification model ***NarationDep*** to automatically detect depressed users based on a hierarchical Attention network is proposed, which exploits data from Twitter. ***NarationDep*** has a component called Hierarchical Attention Based Clustering Network (HACN), which considers the hierarchical structure of user cluster tweets (cluster, tweets and words) and contains an attention mechanism that can find the most crucial cluster that represents the narrative explanation in a user document.

- Examined the dynamics of community depression in NSW. And based on the findings, we demonstrated that people's levels of depression increased as a result of the COVID-19 epidemic. The sudden changes in COVID-19 confirmed cases also had an impact on people's levels of depression. To build depression classification models, a novel approach based on multi-modal features from tweets and Term Frequency-Inverse Document Frequency (TF-IDF) is proposed. The depression cues from topical, emotional, and domain-specific viewpoints are all captured by multi-modal features.

- We have provided a large COVID-19 dataset that may be used to investigate depression on social media in order to understand how COVID-19 has affected people's depression. We also developed a novel method based on Hierarchical Convolutional Neural Network (HCN), which extracts precise and pertinent content from user historical postings, to model the tweets of depressed and non-depressed users before and after the onset of the COVID-19 epidemic. HCN considers the user tweets' hierarchical structure and has an attention technique that can find the key terms and tweets in a user document while also taking the surrounding context into consideration. Our novel method can identify depressed users who are present within the COVID-19 time period. According to our findings using

benchmark datasets, the COVID-19 pandemic caused a large number of previously healthy individuals to develop depression.

8.2 Future work

Although the solutions proposed in this thesis addressed some research problems in misbehavior analysis in social networks, there are still some problems needed to be researched in the future, e.g., lack of ground-truth labels, difficulty diagnosing depression etc. The specific directions for future research are as follows:

- As a future work, we can use short-text topic modeling as a feature extractor to detect mental illness, with the help of word embedding that is learned and trained on a large corpus of depression-related tweets. It is because the word embedding can bring supplemental semantics to the short texts to overcome their lacking of rich contexts. New features can also be tested, and new algorithms can be used and studied for the purpose, how can we reduce the time needed for the prediction.
- Besides the users' tweets, we will analyze user behaviors related to depression during the COVID pandemic, such as social engagement and social interaction with others. This would provide the model with more contextual information and allow us to concentrate on a task where our model not only detects depression but also automatically gives a possible diagnosis. Moreover, we will aim to detect a user's loneliness during the pandemic, which is one mental illness that has never been before in this field. Loneliness is considered one of the early depression symptoms; therefore, its detection will help in early depression detection.
- A model that jointly summarizes user posts and detect mental illness. User tweets summarization and user tweets classification are essential tasks to help detect depressed users through social media. The tweet summarization aims to generate a summary with the significant tweets related to depression from user posts history. User tweets classification is to classify whether a tweet is a depression or not, which is considered a binary classification task. Therefore, we will propose a novel dual model that jointly improves these two tasks' performance to efficiently leverage the shared depression information in both User tweets summarization and user tweets classification tasks.

BIBLIOGRAPHY

[1] A. ACAR AND Y. MURAKI, *Twitter for crisis communication: lessons learned from japan's tsunami disaster*, International Journal of Web Based Communities, 7 (2011), pp. 392–402.

[2] M. E. ADDIS AND N. S. JACOBSON, *Reasons for depression and the process and outcome of cognitive–behavioral psychotherapies.*, Journal of consulting and clinical psychology, 64 (1996), p. 1417.

[3] N. AL ASAD, M. A. M. PRANTO, S. AFREEN, AND M. M. ISLAM, *Depression detection by analyzing social media posts of user*, in SPICSCON, IEEE, 2019, pp. 13–17.

[4] T. AL HANAI, M. M. GHASSEMI, AND J. R. GLASS, *Detecting depression with audio/text sequence modeling of interviews.*, in Interspeech, 2018, pp. 1716–1720.

[5] A. E. ALADAĞ, S. MUDERRISOGLU, N. B. AKBAS, O. ZAHMACIOGLU, AND H. O. BINGOL, *Detecting suicidal ideation on forums: proof-of-concept study*, JMIR, 20 (2018), p. e215.

[6] S. ALMOUZINI, M. KHEMAKHEM, AND A. ALAGEEL, *Detecting arabic depressed users from twitter data*, Procedia Computer Science, 163 (2019), pp. 257–265.

[7] M. E. ARAGÓN, A. P. LÓPEZ-MONROY, L. C. GONZÁLEZ-GURROLA, AND M. MONTES, *Detecting depression in social media using fine-grained emotions*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 1481–1486.

[8] A. P. ASSOCIATION ET AL., *Diagnostic and statistical manual of mental disorders (DSM-5®)*, vol. 49, American Psychiatric Pub, 2013.

[9] D. BAHDANAU, K. CHO, AND Y. BENGIO, *Neural machine translation by jointly learning to align and translate*, arXiv preprint arXiv:1409.0473, (2014).

[10] R. K. BAKSHI, N. KAUR, R. KAUR, AND G. KAUR, *Opinion mining and sentiment analysis*, in 2016 3rd international conference on computing for sustainable global development (INDIACom), IEEE, 2016, pp. 452–455.

[11] G. BARKUR, VIBHA, AND G. B. KAMATH, *Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from india*, Asian Journal of Psychiatry, (2020).

[12] K. C. BATHINA, M. T. THIJ, L. LORENZO-LUACES, L. A. RUTTER, AND J. BOLLEN, *Depressed individuals express more distorted thinking on social media*, arXiv preprint arXiv:2002.02800, (2020).

[13] M. BHAT, M. QADRI, N.-U.-A. BEG, M. KUNDROO, N. AHANGER, AND B. AGARWAL, *Sentiment analysis of social media response on the covid19 outbreak*, Brain, Behavior, and Immunity, (2020).

[14] S. BHATTACHARJEE, L. GOLDSTONE, N. VADIEI, J. K. LEE, AND W. J. BURKE, *Depression screening patterns, predictors, and trends among adults without a depression diagnosis in ambulatory settings in the united states*, Psychiatric services, 69 (2018), pp. 1098–1100.

[15] D. M. BLEI, A. Y. NG, AND M. I. JORDAN, *Latent dirichlet allocation*, Journal of machine Learning research, 3 (2003), pp. 993–1022.

[16] P. BOJANOWSKI, E. GRAVE, A. JOULIN, AND T. MIKOLOV, *Enriching word vectors with subword information*, TACL, 5 (2017), pp. 135–146.

[17] M. M. BRADLEY AND P. J. LANG, *Affective norms for english words (anew): Instruction manual and affective ratings*, tech. rep., Technical report C-1, the center for research in psychophysiology ‚Ä¶, 1999.

[18] M. BROERSMA AND T. GRAHAM, *Social media as beat: Tweets as a news source during the 2010 british and dutch elections*, journalism practice, 6 (2012), pp. 403–419.

[19] S. BUCCI, M. SCHWANNAUER, AND N. BERRY, *The digital revolution and its impact on mental health care*, Psychology and Psychotherapy: Theory, Research and Practice, 92 (2019), pp. 277–297.

[20] R. CATIPON AND O. SAVAS, *A different story: How conservative narratives diverge between twitter and parler.*, in Text2Story@ ECIR, 2021, pp. 15–22.

[21] N. CHAMBERS AND D. JURAFSKY, *Unsupervised learning of narrative event chains*, in Proceedings of ACL-08: HLT, 2008, pp. 789–797.

[22] S. CHANCELLOR AND M. DE CHOUDHURY, *Methods in predictive techniques for mental health status on social media: a critical review*, NPJ digital medicine, 3 (2020), pp. 1–11.

[23] H. CHEN, Y. LI, X. SUN, G. XU, AND H. YIN, *Temporal meta-path guided explainable recommendation*, in Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 2021, pp. 1056–1064.

[24] Y. CHEN, B. ZHOU, W. ZHANG, W. GONG, AND G. SUN, *Sentiment analysis based on deep learning and its application in screening for perinatal depression*, in DSC, IEEE, 2018, pp. 451–456.

[25] C. Y. CHIU, H. Y. LANE, J. L. KOH, AND A. L. CHEN, *Multimodal depression detection on instagram considering time interval of posts*, Journal of Intelligent Information Systems, 56 (2020), pp. 1–23.

[26] K. CHO, B. VAN MERRIËNBOER, C. GULCEHRE, D. BAHDANAU, F. BOUGARES, H. SCHWENK, AND Y. BENGIO, *Learning phrase representations using RNN encoder–decoder for statistical machine translation*, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, Oct. 2014, Association for Computational Linguistics, pp. 1724–1734.

[27] E. P. H. CHOI, B. P. H. HUI, AND E. Y. F. WAN, *Depression and anxiety in hong kong during covid-19*, International journal of environmental research and public health, 17 (2020), p. 3740.

[28] J. CHUNG, C. GULCEHRE, K. CHO, AND Y. BENGIO, *Empirical evaluation of gated recurrent neural networks on sequence modeling*, arXiv preprint arXiv:1412.3555, (2014).

[29] D. CONG, Y. ZHAO, B. QIN, Y. HAN, M. ZHANG, A. LIU, AND N. CHEN, *Hierarchical attention based neural network for explainable recommendation*, in

Proceedings of the 2019 on International Conference on Multimedia Retrieval, 2019, pp. 373–381.

[30] Q. CONG, Z. FENG, F. LI, Y. XIANG, G. RAO, AND C. TAO, *Xa-bilstm: a deep learning approach for depression detection in imbalanced data*, in 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2018, pp. 1624–1627.

[31] M. CONWAY AND D. O,ÄôCONNOR, *Social media, big data, and mental health: current advances and ethical implications*, Current opinion in psychology, 9 (2016), pp. 77–82.

[32] G. COPPERSMITH, M. DREDZE, AND C. HARMAN, *Quantifying mental health signals in twitter*, in CLCP, 2014, pp. 51–60.

[33] ——, *Quantifying mental health signals in twitter*, in CLCP, June 2014, pp. 51–60.

[34] G. COPPERSMITH, M. DREDZE, C. HARMAN, AND K. HOLLINGSHEAD, *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, 2015, pp. 1–10.

[35] D. CUCINOTTA AND M. VANELLI, *Who declares covid-19 a pandemic*, Acta Bio Medica: Atenei Parmensis, 91 (2020), p. 157.

[36] M. DE CHOUDHURY, S. COUNTS, AND E. HORVITZ, *Predicting postpartum changes in emotion and behavior via social media*, in SIGCHI, 2013, pp. 3267–3276.

[37] ——, *Social media as a measurement tool of depression in populations*, in Proceedings of the 5th annual ACM web science conference, 2013, pp. 47–56.

[38] M. DE CHOUDHURY, S. COUNTS, E. J. HORVITZ, AND A. HOFF, *Characterizing and predicting postpartum depression from shared facebook data*, in CSCWC, 2014, pp. 626–638.

[39] M. DESHPANDE AND V. RAO, *Depression detection using emotion artificial intelligence*, in 2017 International Conference on Intelligent Sustainable Systems (ICISS), Dec 2017, pp. 858–862.

[40] M. DESHPANDE AND V. RAO, *Depression detection using emotion artificial intelligence*, in ICISS, IEEE, 2017, pp. 858–862.

[41] J. DEVLIN, M.-W. CHANG, K. LEE, AND K. TOUTANOVA, *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv preprint arXiv:1810.04805, (2018).

[42] Y. DING, X. CHEN, Q. FU, AND S. ZHONG, *A depression recognition method for college students using deep integrated support vector algorithm*, IEEE Access, 8 (2020), pp. 75616–75629.

[43] F. DOSHI-VELEZ AND B. KIM, *Towards a rigorous science of interpretable machine learning, corr abs/1702.08608*, arXiv preprint arXiv:1702.08608, (2017).

[44] M. DU, N. LIU, AND X. HU, *Techniques for interpretable machine learning*, Communications of the ACM, 63 (2019), pp. 68–77.

[45] P. S. EARLE, D. C. BOWDEN, AND M. GUY, *Twitter earthquake detection: earthquake monitoring in a social world*, Annals of Geophysics, 54 (2012).

[46] F. EDITION ET AL., *Diagnostic and statistical manual of mental disorders*, Am Psychiatric Assoc, 21 (2013).

[47] J. C. EICHSTAEDT, R. J. SMITH, R. M. MERCHANT, L. H. UNGAR, P. CRUTCHLEY, D. PREOŢIUC-PIETRO, D. A. ASCH, AND H. A. SCHWARTZ, *Facebook language predicts depression in medical records*, Proceedings of the National Academy of Sciences, 115 (2018), pp. 11203–11208.

[48] A. ESSIEN, I. PETROUNIAS, P. SAMPAIO, AND S. SAMPAIO, *A deep-learning model for urban traffic flow prediction with traffic events mined from twitter*, World Wide Web, 24 (2021), pp. 1345–1368.

[49] N. FARRUQUE, O. ZAIANE, AND R. GOEBEL, *Augmenting semantic representation of depressive language: From forums to microblogs*, in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2019, pp. 359–375.

[50] M. FARUQUI, J. DODGE, S. K. JAUHAR, C. DYER, E. HOVY, AND N. A. SMITH, *Retrofitting word vectors to semantic lexicons*, arXiv preprint arXiv:1411.4166, (2014).

[51] I. FATIMA, H. MUKHTAR, H. F. AHMAD, AND K. RAJPOOT, *Analysis of user-generated content from online social communities to characterise and predict depression degree*, Journal of Information Science, 44 (2018), pp. 683–695.

[52] S. GALEA, R. M. MERCHANT, AND N. LURIE, *The mental health consequences of covid-19 and physical distancing: the need for prevention and early intervention*, JAMA internal medicine, 180 (2020), pp. 817–818.

[53] J. GERACI, P. WILANSKY, V. DE LUCA, A. ROY, J. L. KENNEDY, AND J. STRAUSS, *Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression*, Evidence-based mental health, 20 (2017), pp. 83–87.

[54] S. GHODRATNAMA, A. BEHESHTI, M. ZAKERSHAHRAK, AND F. SOBHANMANESH, *Intelligent narrative summaries: From indicative to informative summarization*, Big Data Research, 26 (2021), p. 100257.

[55] J. GIBBONS, R. MALOUF, B. SPITZBERG, L. MARTINEZ, B. APPLEYARD, C. THOMPSON, A. NARA, AND M.-H. TSOU, *Twitter-based measures of neighborhood sentiment as predictors of residential population health*, PLoS ONE, 14 (2019).

[56] Y. M. GOH AND C. UBEYNARAYANA, *Construction accident narrative classification: An evaluation of text mining techniques*, Accident Analysis & Prevention, 108 (2017), pp. 122–130.

[57] K. GOYAL, P. CHAUHAN, K. CHHIKARA, P. GUPTA, AND M. P. SINGH, *Fear of covid 2019: First suicidal case in india!*, (2020).

[58] T. GUI, L. ZHU, Q. ZHANG, M. PENG, X. ZHOU, K. DING, AND Z. CHEN, *Cooperative multimodal approach to depression detection in twitter*, in AAAI, vol. 33, 2019, pp. 110–117.

[59] I. GUSEV, *Dataset for automatic summarization of russian news*, in AINL, Springer, 2020, pp. 122–134.

[60] C. HAWN, *Take two aspirin and tweet me in the morning: how twitter, facebook, and other social media are reshaping health care*, Health affairs, 28 (2009), pp. 361–368.

[61] K. HE, X. ZHANG, S. REN, AND J. SUN, *Deep residual learning for image recognition*, in ICCV, 2016, pp. 770–778.

[62] S. HOCHREITER AND J. SCHMIDHUBER, *Long short-term memory*, Neural computation, 9 (1997), pp. 1735–1780.

[63] M. HOLMSTROM ET AL., *The narrative and social media*, Defence Strategic Communications, 1 (2015), pp. 118–132.

[64] K. HORECKI AND J. MAZURKIEWICZ, *Natural language processing methods used for automatic prediction mechanism of related phenomenon*, Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science), 9120 (2015), pp. 13–24.

[65] ——, *Natural language processing methods used for automatic prediction mechanism of related phenomenon*, in International Conference on Artificial Intelligence and Soft Computing, Springer, 2015, pp. 13–24.

[66] J. HOWARD AND S. RUDER, *Universal language model fine-tuning for text classification*, arXiv preprint arXiv:1801.06146, (2018).

[67] Q. HU, A. LI, F. HENG, J. LI, AND T. ZHU, *Predicting depression of social media user on different observation windows*, in 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), vol. 1, IEEE, 2015, pp. 361–364.

[68] J. HUA AND R. SHAW, *Corona virus (covid-19) "infodemic" and emerging issues through a data lens: The case of china*, International journal of environmental research and public health, 17 (2020), p. 2309.

[69] Y. HUANG AND N. ZHAO, *Generalized anxiety disorder, depressive symptoms and sleep quality during covid-19 outbreak in china: a web-based cross-sectional survey*, Psychiatry research, 288 (2020), p. 112954.

[70] M. N. HUSSAIN, H. AL RUBAYE, K. K. BANDELI, AND N. AGARWAL, *Stories from blogs: Computational extraction and visualization of narratives.*, in Text2Story@ ECIR, 2021, pp. 33–40.

[71] A. HUSSEINI ORABI, P. BUDDHITHA, M. HUSSEINI ORABI, AND D. INKPEN, *Deep learning for depression detection of twitter users*, in CLCP, New Orleans, LA, June 2018, pp. 88–97.

[72] K. JAIDKA, S. GIORGI, H. A. SCHWARTZ, M. L. KERN, L. H. UNGAR, AND J. C. EICHSTAEDT, *Estimating geographic subjective well-being from twitter: A comparison of dictionary and data-driven language methods*, Proceedings of the National Academy of Sciences, 117 (2020), pp. 10165–10171.

[73] S. L. JAMES, D. ABATE, K. H. ABATE, S. M. ABAY, C. ABBAFATI, N. ABBASI, H. ABBASTABAR, F. ABD-ALLAH, J. ABDELA, A. ABDELALIM, ET AL., *Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017*, The Lancet, 392 (2018), pp. 1789–1858.

[74] L. KANG, S. MA, M. CHEN, J. YANG, Y. WANG, R. LI, L. YAO, H. BAI, Z. CAI, B. X. YANG, ET AL., *Impact on mental health and perceptions of psychological care among medical and nursing staff in wuhan during the 2019 novel coronavirus disease outbreak: A cross-sectional study*, Brain, behavior, and immunity, 87 (2020), pp. 11–17.

[75] C. KARMEN, R. C. HSIUNG, AND T. WETTER, *Screening internet forum participants for depression symptoms by assembling and enhancing multiple nlp methods*, Computer methods and programs in biomedicine, 120 (2015), pp. 27–36.

[76] J. KIM, J. LEE, E. PARK, AND J. HAN, *A deep learning model for detecting mental illness from user content on social media*, Scientific reports, 10 (2020), pp. 1–6.

[77] Y. KIM, *Convolutional neural networks for sentence classification*, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, Oct. 2014, Association for Computational Linguistics, pp. 1746–1751.

[78] D. P. KINGMA AND J. BA, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980, (2014).

[79] D. P. KINGMA AND J. BA, *Adam: A method for stochastic optimization*, CoRR, abs/1412.6980 (2015).

[80] K. KIRA, L. A. RENDELL, ET AL., *The feature selection problem: Traditional methods and a new algorithm*, in Aaai, vol. 2, 1992, pp. 129–134.

[81] A. L. KLEPPANG, I. HARTZ, M. THURSTON, AND C. HAGQUIST, *The association between physical activity and symptoms of depression in different contexts–a cross-sectional study of norwegian adolescents*, BMC public health, 18 (2018), pp. 1–12.

[82] E. J. KLOSTERMAN, *Text mining of patient demographics and diagnoses from psychiatric assessments*, (2014).

[83] D. KOH, *Covid-19 lockdowns throughout the world*, Occupational Medicine, 70 (2020), pp. 322–322.

[84] A. KOLLIAKOU, I. BAKOLIS, D. CHANDRAN, L. DERCZYNSKI, N. WERBELOFF, D. P. OSBORN, K. BONTCHEVA, AND R. STEWART, *Mental health-related conversations on social media and crisis episodes: a time-series regression analysis*, Scientific reports, 10 (2020), pp. 1–7.

[85] A. KUMAR, A. SHARMA, AND A. ARORA, *Anxious depression prediction in real-time social data*, Available at SSRN 3383359, (2019).

[86] R. LEBRET AND R. COLLOBERT, *Rehabilitation of count-based models for word vector representations*, in International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2015, pp. 417–429.

[87] A. LEIS, F. RONZANO, M. A. MAYER, L. I. FURLONG, AND F. SANZ, *Detecting signs of depression in tweets in spanish: Behavioral and linguistic analysis*, JMIR, 21 (2019).

[88] M. LEWIS, Y. LIU, N. GOYAL, M. GHAZVININEJAD, A. MOHAMED, O. LEVY, V. STOYANOV, AND L. ZETTLEMOYER, *Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*, arXiv preprint arXiv:1910.13461, (2019).

[89] I. LI, Y. LI, T. LI, S. ALVAREZ-NAPAGAO, D. GARCIA-GASULLA, AND T. SUZUMURA, *What are we depressed about when we talk about COVID19: Mental health analysis on tweets using natural language processing*, arXiv:2004.10899 [cs], (2020).

[90] C. LIN, P. HU, H. SU, S. LI, J. MEI, J. ZHOU, AND H. LEUNG, *Sensemood: Depression detection on social media*, in Proceedings of the 2020 International Conference on Multimedia Retrieval, 2020, pp. 407–411.

[91] H. LIN, J. JIA, J. QIU, Y. ZHANG, G. SHEN, L. XIE, J. TANG, L. FENG, AND T. CHUA, *Detecting stress based on social interactions in social networks*, IEEE Transactions on Knowledge and Data Engineering, 29 (2017), pp. 1820–1833.

[92] N. LIU, M. DU, AND X. HU, *Representation interpretation with spatial encoding and multimodal analytics*, in Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 2019, pp. 60–68.

[93] N. LIU, H. YANG, AND X. HU, *Adversarial detection with model interpretation*, in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 1803–1811.

[94] P. J. LIU, M. SALEH, E. POT, B. GOODRICH, R. SEPASSI, L. KAISER, AND N. SHAZEER, *Generating wikipedia by summarizing long sequences*, arXiv preprint arXiv:1801.10198, (2018).

[95] Y. LIU AND M. LAPATA, *Hierarchical transformers for multi-document summarization*, arXiv preprint arXiv:1905.13164, (2019).

[96] Y. LIU, M. OTT, N. GOYAL, J. DU, M. JOSHI, D. CHEN, O. LEVY, M. LEWIS, L. ZETTLEMOYER, AND V. STOYANOV, *Roberta: A robustly optimized bert pretraining approach*, arXiv preprint arXiv:1907.11692, (2019).

[97] D. E. LOSADA, F. CRESTANI, AND J. PARAPAR, *erisk 2017: Clef lab on early risk prediction on the internet: experimental foundations*, in International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2017, pp. 346–360.

[98] ——, *Overview of erisk: early risk prediction on the internet*, in International conference of the cross-language evaluation forum for european languages, Springer, 2018, pp. 343–361.

[99] D. M. LOW, L. RUMKER, T. TALKAR, J. TOROUS, G. CECCHI, AND S. S. GHOSH, *Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study*, Journal of medical Internet research, 22 (2020), p. e22635.

[100] W. LU, *Treatment for adolescent depression: national patterns, temporal trends, and factors related to service use across settings*, Journal of Adolescent Health, 67 (2020), pp. 401–408.

[101] N. F. LUND, S. A. COHEN, AND C. SCARLES, *The power of social media storytelling in destination branding*, Journal of destination marketing & management, 8 (2018), pp. 271–280.

[102] D. LYU, Z. WANG, Y. DU, R. K. MARJERISON, AND R. CHEN, *Using social media content to identify mental health problems: The case of# depression in sina weibo*, Review of Integrative Business and Economics Research, 9 (2020), pp. 448–464.

[103] C. MANNING AND H. SCHUTZE, *Foundations of statistical natural language processing*, MIT press, 1999.

[104] C. MCCLELLAN, M. M. ALI, R. MUTTER, L. KROUTIL, AND J. LANDWEHR, *Using social media to monitor mental health discussions- evidence from twitter*, JAMIA, 24 (2017), pp. 496–502.

[105] J. M. METZL AND K. T. MACLEISH, *Mental illness, mass shootings, and the politics of american firearms*, American journal of public health, 105 (2015), pp. 240–249.

[106] T. MIKOLOV, K. CHEN, G. CORRADO, J. DEAN, L. SUTSKEVER, AND G. ZWEIG, *word2vec*, URL https://code. google. com/p/word2vec, (2013).

[107] T. MIKOLOV, E. GRAVE, P. BOJANOWSKI, C. PUHRSCH, AND A. JOULIN, *Advances in pre-training distributed word representations*, arXiv preprint arXiv:1712.09405, (2017).

[108] T. MIKOLOV, I. SUTSKEVER, K. CHEN, G. S. CORRADO, AND J. DEAN, *Distributed representations of words and phrases and their compositionality*, in NIPS, 2013, pp. 3111–3119.

[109] D. MILLER, *Leveraging bert for extractive text summarization on lectures*, arXiv preprint arXiv:1906.04165, (2019).

[110] L. J. MILLER, *Postpartum depression*, Jama, 287 (2002), pp. 762–765.

[111] C. MISPERCEPTIONS, *Mass shootings and mental illness*, (2016).

[112] N. MONTEMURRO, *The emotional impact of COVID-19: From medical staff to common people*, Brain, Behavior, and Immunity, (2020).

[113] L. MOSSBURGER, F. WENDE, K. BRINKMANN, AND T. SCHMIDT, *Exploring online depression forums via text mining: a comparison of reddit and a curated online forum*, in Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task, 2021, pp. 70–81.

[114] M. NADEEM, *Identifying depression on twitter*, arXiv preprint arXiv:1607.07384, (2016).

[115] R. NALLAPATI, B. ZHOU, C. GULCEHRE, B. XIANG, ET AL., *Abstractive text summarization using sequence-to-sequence rnns and beyond*, arXiv preprint arXiv:1602.06023, (2016).

[116] L. NEUHAUSER AND G. L. KREPS, *Rethinking communication in the e-health era*, Journal of Health Psychology, 8 (2003), pp. 7–23.

[117] A. Y. NG AND M. I. JORDAN, *On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes*, in Advances in neural information processing systems, 2002, pp. 841–848.

[118] E. NG, *The pandemic of hate is giving covid-19 a helping hand*, The American journal of tropical medicine and hygiene, 102 (2020), p. 1158.

[119] T. H. NGUYEN AND R. GRISHMAN, *Relation extraction: Perspective from convolutional neural networks*, in VSM for NLP, Denver, Colorado, June 2015, pp. 39–48.

[120] H. NI, S. WANG, AND P. CHENG, *A hybrid approach for stock trend prediction based on tweets embedding and historical prices*, World Wide Web, 24 (2021), pp. 849–868.

[121] M. Y. NI, L. YANG, C. M. LEUNG, N. LI, X. I. YAO, Y. WANG, G. M. LEUNG, B. J. COWLING, AND Q. LIAO, *Mental health, risk factors, and social media use during the covid-19 epidemic and cordon sanitaire among the community and health professionals in wuhan, china: Cross-sectional survey*, JMIR, 7 (2020), p. e19009.

[122] P. K. NOVAK, J. SMAILOVIĆ, B. SLUBAN, AND I. MOZETIČ, *Sentiment of emojis*, PloS one, 10 (2015), p. e0144296.

[123] E. A. O'CONNOR, E. P. WHITLOCK, B. GAYNES, AND T. L. BEIL, *Screening for depression in adults and older adults in primary care: An updated systematic review*, Evidence Syntheses 75, Agency for Healthcare Research and Quality (US), December 2009.

[124] N. OFEK, G. KATZ, B. SHAPIRA, AND Y. BAR-ZEV, *Sentiment analysis in transcribed utterances*, in Pacific-Asia conference on knowledge discovery and data mining, Springer, 2015, pp. 27–38.

[125] A. H. ORABI, P. BUDDHITHA, M. H. ORABI, AND D. INKPEN, *Deep learning for depression detection of twitter users*, in Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, 2018, pp. 88–97.

[126] W. H. ORGANIZATION, *Depression*.

[127] R. PAGE, *The narrative dimensions of social media storytelling*, The handbook of narrative analysis, (2015), pp. 329–347.

[128] B. PANG, L. LEE, AND S. VAITHYANATHAN, *Thumbs up? sentiment classification using machine learning techniques*, arXiv preprint cs/0205070, (2002).

[129] M. PARK, C. CHA, AND M. CHA, *Depressive moods of users portrayed in twitter*, in Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD), vol. 2012, 2012, pp. 1–8.

[130] J. D. G. PAULE, Y. SUN, AND Y. MOSHFEGHI, *On fine-grained geolocalisation of tweets and real-time traffic incident detection*, IPM, 56 (2019), pp. 1119–1132.

[131] F. C. PAYTON, L. K. YARGER, AND A. T. PINTER, *Text mining mental health reports for issues impacting today,Äôs college students: qualitative study*, JMIR mental health, 5 (2018), p. e10032.

[132] F. PEDREGOSA, G. VAROQUAUX, A. GRAMFORT, V. MICHEL, B. THIRION, O. GRISEL, M. BLONDEL, P. PRETTENHOFER, R. WEISS, V. DUBOURG, ET AL., *Scikit-learn: Machine learning in python*, the Journal of machine Learning research, 12 (2011), pp. 2825–2830.

[133] Z. PENG, Q. HU, AND J. DANG, *Multi-kernel svm based depression recognition using social media data*, IJMLC, 10 (2019), pp. 43–57.

[134] J. PENNINGTON, R. SOCHER, AND C. D. MANNING, *Glove: Global vectors for word representation*, in EMNLP, 2014, pp. 1532–1543.

[135] B. W. PENNINX, Y. MILANESCHI, F. LAMERS, AND N. VOGELZANGS, *Understanding the somatic consequences of depression: biological mechanisms and the role of depression symptom profile*, BMC Medicine, 11 (2013), p. 129.

[136] G. PENNYCOOK, J. MCPHETRES, Y. ZHANG, J. G. LU, AND D. G. RAND, *Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention*, Psychological science, 31 (2020), pp. 770–780.

[137] K. W. PRIER, M. S. SMITH, C. GIRAUD-CARRIER, AND C. L. HANSON, *Identifying health-related topics on twitter*, in International conference on social computing, behavioral-cultural modeling, and prediction, Springer, 2011, pp. 18–25.

[138] A. RAMESH, M. PAVLOV, G. GOH, S. GRAY, C. VOSS, A. RADFORD, M. CHEN, AND I. SUTSKEVER, *Zero-shot text-to-image generation*, in International Conference on Machine Learning, PMLR, 2021, pp. 8821–8831.

[139] D. RAMÍREZ-CIFUENTES, M. MAYANS, AND A. FREIRE, *Early risk detection of anorexia on social media*, in International Conference on Internet Science, Springer, 2018, pp. 3–14.

[140] G. RAO, Y. ZHANG, L. ZHANG, Q. CONG, AND Z. FENG, *Mgl-cnn: A hierarchical posts representations model for identifying depressed individuals in online forums*, IEEE Access, 8 (2020), pp. 32395–32403.

[141] C. S. A. RAZAK, M. A. ZULKARNAIN, S. H. A. HAMID, N. B. ANUAR, M. Z. JALI, AND H. MEON, *Tweep: A system development to detect depression in twitter posts*, in Computational Science and Technology, R. Alfred, Y. Lim, H. Haviluddin, and C. K. On, eds., LNEE, Springer, 2020, pp. 543–552.

[142] A. G. REECE, A. J. REAGAN, K. L. M. LIX, P. S. DODDS, C. M. DANFORTH, AND E. J. LANGER, *Forecasting the onset and course of mental illness with twitter data*, Scientific Reports, 7 (2017), p. 13006.

[143] N. REIMERS AND I. GUREVYCH, *Sentence-bert: Sentence embeddings using siamese bert-networks*, arXiv preprint arXiv:1908.10084, (2019).

[144] P. RESNIK, W. ARMSTRONG, L. CLAUDINO, T. NGUYEN, V.-A. NGUYEN, AND J. BOYD-GRABER, *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, Denver, Colorado, June 5 2015, Association for Computational Linguistics, pp. 99–107.

[145] E. A. RÍSSOLA, M. ALIANNEJADI, AND F. CRESTANI, *Beyond modelling: Understanding mental disorders in online social media*, in European Conference on Information Retrieval, Springer, 2020, pp. 296–310.

[146] S. RODRIGUES, B. BOKHOUR, N. MUELLER, N. DELL, P. E. OSEI-BONSU, S. ZHAO, M. GLICKMAN, S. V. EISEN, AND A. R. ELWY, *Impact of stigma on veteran treatment seeking for depression*, AJPR, 17 (2014), pp. 128–146.

[147] J. P. ROGERS, E. CHESNEY, D. OLIVER, T. A. POLLAK, P. MCGUIRE, P. FUSAR-POLI, M. S. ZANDI, G. LEWIS, AND A. S. DAVID, *Psychiatric and neuropsychiatric presentations associated with severe coronavirus infections: a systematic review and meta-analysis with comparison to the COVID-19 pandemic*, The Lancet Psychiatry, (2020).

[148] R. ROSA, T. MUSIL, O. DUŠEK, D. JURKO, P. SCHMIDTOVÁ, D. MAREČEK, O. BOJAR, T. KOCMI, D. HRBEK, D. KOŠŤÁK, ET AL., *Theaitre 1.0: Interactive generation of theatre play scripts*, arXiv preprint arXiv:2102.08892, (2021).

[149] A. S. ROSS AND D. J. RIVERS, *Discursive deflection: Accusation of ‚Äúfake news‚Äù and the spread of mis-and disinformation in the tweets of president trump*, Social Media+ Society, 4 (2018), p. 2056305118776010.

[150] R. ROSSI, V. SOCCI, D. TALEVI, S. MENSI, C. NIOLU, F. PACITTI, A. DI MARCO, A. ROSSI, A. SIRACUSANO, AND G. DI LORENZO, *Covid-19 pandemic and lockdown measures impact on mental health among the general population in italy. front psychiatry. 2020; 11: 790*, 2020.

[151] Z. SAEED, R. A. ABBASI, I. RAZZAK, O. MAQBOOL, A. SADAF, AND G. XU, *Enhanced heartbeat graph for emerging event detection on twitter using time series networks*, ESA, 136 (2019), pp. 115–132.

[152] J. SANKARANARAYANAN, H. SAMET, B. E. TEITLER, M. D. LIEBERMAN, AND J. SPERLING, *Twitterstand: news in tweets*, in SIGSPATIAL, 2009, pp. 42–51.

[153] D. SCANFELD, V. SCANFELD, AND E. L. LARSON, *Dissemination of health information through social networks: Twitter and antibiotics*, American journal of infection control, 38 (2010), pp. 182–188.

[154] B.-W. SEMO AND S. M. FRISSA, *The mental health impact of the covid-19 pandemic: implications for sub-saharan africa*, Psychology Research and Behavior Management, 13 (2020), p. 713.

[155] A. B. SHATTE, D. M. HUTCHINSON, M. FULLER-TYSZKIEWICZ, AND S. J. TEAGUE, *Social media markers to identify fathers at risk of postpartum depression: a machine learning approach*, Cyberpsychology, Behavior, and Social Networking, 23 (2020), pp. 611–618.

[156] G. SHEN, J. JIA, L. NIE, F. FENG, C. ZHANG, T. HU, T.-S. CHUA, AND W. ZHU, *Depression detection via harvesting social media: A multimodal dictionary learning solution.*, in IJCAI, 2017, pp. 3838–3844.

[157] T. SHEN, J. JIA, G. SHEN, F. FENG, X. HE, H. LUAN, J. TANG, T. TIROPANIS, T. S. CHUA, AND W. HALL, *Cross-domain depression detection via harvesting social media*, in Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI 2018, vol. 2018-July, International Joint Conferences on Artificial Intelligence, July 2018, pp. 1611–1617.

[158] T. SHEN, J. JIA, G. SHEN, F. FENG, X. HE, H. LUAN, J. TANG, T. TIROPANIS, T. S. CHUA, AND W. HALL, *Cross-domain depression detection via harvesting social media*, in IJCAI, IJCAI, 2018.

[159] X. SHEN, X. ZOU, X. ZHONG, J. YAN, AND L. LI, *Psychological stress of icu nurses in the time of covid-19*, 2020.

[160] B. SHMUELI, *Multi-class metrics made simple, part ii: the f1-score*, Retrieved from Towards Data Science: https://towardsdatascience. com/multi-class-metrics-made-simplepart-ii-the-f1-score-ebe8b2c2ca1, (2019).

[161] A. SHRESTHA, E. SERRA, AND F. SPEZZANO, *Multi-modal social and psycholinguistic embedding via recurrent neural networks to identify depressed users in online forums.*, NetMAHIB, 9 (2020), p. 22.

[162] K. SIBELIUS, *Increasing access to mental health services*.

http://www.whitehouse.gov/blog/2013/04/10/increasingaccess-mental-health-services. 2013.

[163] T. SIMMS, C. RAMSTEDT, M. RICH, M. RICHARDS, T. MARTINEZ, AND C. GIRAUD-CARRIER, *Detecting cognitive distortions through machine learning text analytics*, in 2017 IEEE international conference on healthcare informatics (ICHI), IEEE, 2017, pp. 508–512.

[164] R. SINGH, J. DU, Y. ZHANG, H. WANG, Y. MIAO, O. A. SIANAKI, AND A. ULHAQ, *A framework for early detection of antisocial behavior on twitter using natural language processing*, in Conference on Complex, Intelligent, and Software Intensive Systems, Springer, 2019, pp. 484–495.

[165] S. SINGH, D. ROY, K. SINHA, S. PARVEEN, G. SHARMA, AND G. JOSHI, *Impact of covid-19 and lockdown on mental health of children and adolescents: A narrative review with recommendations*, Psychiatry research, 293 (2020), p. 113429.

[166] S. SRIVASTAVA, S. CHATURVEDI, AND T. MITCHELL, *Inferring interpersonal relations in narrative summaries*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30, 2016.

[167] D. STEPHENS AND M. DIESING, *A comparison of supervised classification methods for the prediction of substrate type using multibeam acoustic and legacy grain-size data*, PloS one, 9 (2014), p. e93950.

[168] Q. SUN, Q. YUE, F. ZHU, AND K. SHU, *The identification research of bipolar disorder based on cnn.[j]*, in Journal of Physics, vol. 1168, 2019.

[169] V. SURVEILLANCES, *The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (covid-19),Äîchina, 2020*, China CDC Weekly, 2 (2020), pp. 113–122.

[170] K. TAGO, K. TAKAGI, S. KASUYA, AND Q. JIN, *Analyzing influence of emotional tweets on user relationships using naive bayes and dependency parsing*, World Wide Web, 22 (2019), pp. 1263–1278.

[171] J. TAN, X. WAN, AND J. XIAO, *Abstractive document summarization with a graph-based attentional neural model*, in ACL, 2017, pp. 1171–1181.

[172] O. TAS AND F. KIYANI, *A survey automatic text summarization*, PressAcademia Procedia, 5 (2007), pp. 205–213.

[173] Y. W. TEH, M. I. JORDAN, M. J. BEAL, AND D. M. BLEI, *Sharing clusters among related groups: Hierarchical dirichlet processes*, in Advances in neural information processing systems, 2005, pp. 1385–1392.

[174] J. C. TOLENTINO AND S. L. SCHMIDT, *Dsm-5 criteria and depression severity: implications for clinical practice*, Frontiers in psychiatry, 9 (2018), p. 450.

[175] P. TRIRAT AND J.-G. LEE, *Df-tar: A deep fusion network for citywide traffic accident risk prediction with dangerous driving behavior*, in Proceedings of the Web Conference 2021, 2021, pp. 1146–1156.

[176] M. TROTZEK, S. KOITKA, AND C. M. FRIEDRICH, *Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences*, TKDE, 32 (2018), pp. 588–601.

[177] ——, *Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia.*, in CLEF (Working Notes), 2018.

[178] S. TSUGAWA, Y. KIKUCHI, F. KISHINO, K. NAKAJIMA, Y. ITOH, AND H. OHSAKI, *Recognizing depression from twitter activity*, in HFCI, ACM, 2015, pp. 3187–3196.

[179] S. TSUGAWA, Y. MOGI, Y. KIKUCHI, F. KISHINO, K. FUJITA, Y. ITOH, AND H. OHSAKI, *On estimating depressive tendencies of twitter users utilizing their tweet data*, in VR, IEEE, 2013, pp. 1–4.

[180] M. T. TULL, K. A. EDMONDS, K. M. SCAMALDO, J. R. RICHMOND, J. P. ROSE, AND K. L. GRATZ, *Psychological outcomes associated with stay-at-home orders and the perceived impact of covid-19 on daily life*, Psychiatry research, 289 (2020), p. 113098.

[181] P. D. TURNEY, *Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews*, arXiv preprint cs/0212032, (2002).

[182] T. UMEMATSU, A. SANO, S. TAYLOR, AND R. W. PICARD, *Improving students' daily life stress forecasting using lstm neural networks*, in BHI, IEEE, 2019, pp. 1–4.

[183] J. J. VAN BAVEL, K. BAICKER, P. S. BOGGIO, V. CAPRARO, A. CICHOCKA, M. CIKARA, M. J. CROCKETT, A. J. CRUM, K. M. DOUGLAS, J. N. DRUCKMAN, ET AL., *Using social and behavioural science to support covid-19 pandemic response*, Nature human behaviour, 4 (2020), pp. 460–471.

[184] A. VASWANI, N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, Ł. KAISER, AND I. POLOSUKHIN, *Attention is all you need*, in NIPS, 2017, pp. 5998–6008.

[185] N. VEDULA AND S. PARTHASARATHY, *Emotional and linguistic cues of depression from social media*, in Proceedings of the 2017 International Conference on Digital Health, 2017, pp. 127–136.

[186] F. B. VIÉGAS AND M. WATTENBERG, *Timelines tag clouds and the case for vernacular visualization*, interactions, 15 (2008), pp. 49–52.

[187] D. VIGO, G. THORNICROFT, AND R. ATUN, *Estimating the true global burden of mental illness*, The Lancet Psychiatry, 3 (2016), pp. 171–178.

[188] P. VINAYAVEKHIN, S. CHAUDHURY, A. MUNAWAR, D. J. AGRAVANTE, G. DE MAGISTRIS, D. KIMURA, AND R. TACHIBANA, *Focusing on what is relevant: Time-series learning and understanding using attention*, in 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, 2018, pp. 2624–2629.

[189] P. VINOD, S. SAFAR, D. MATHEW, P. VENUGOPAL, L. M. JOLY, AND J. GEORGE, *Fine-tuning the bertsumext model for clinical report summarization*, in INCET, IEEE, 2020, pp. 1–7.

[190] N. N. VO, X. HE, S. LIU, AND G. XU, *Deep learning for decision making and the optimization of socially responsible investments and portfolio*, Decision Support Systems, 124 (2019), p. 113097.

[191] N. N. VO, S. LIU, X. LI, AND G. XU, *Leveraging unstructured call log data for customer churn prediction*, Knowledge-Based Systems, 212 (2021), p. 106586.

[192] F. WANG, J. XU, C. LIU, R. ZHOU, AND P. ZHAO, *On prediction of traffic flows in smart cities: a multitask deep learning based approach*, World Wide Web, 24 (2021), pp. 805–823.

[193] S.-H. WANG, Y. DING, W. ZHAO, Y.-H. HUANG, R. PERKINS, W. ZOU, AND J. J. CHEN, *Text mining for identifying topics in the literatures about adolescent substance use and depression*, BMC public health, 16 (2016), pp. 1–8.

[194] T. WANG, M. BREDE, A. IANNI, AND E. MENTZAKIS, *Detecting and characterizing eating-disorder communities on social media*, in Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, ACM, 2017, pp. 91–100.

[195] W. WANG, N. YANG, F. WEI, B. CHANG, AND M. ZHOU, *Gated self-matching networks for reading comprehension and question answering*, in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 189–198.

[196] X. WANG, S. CHEN, T. LI, W. LI, Y. ZHOU, J. ZHENG, Q. CHEN, J. YAN, AND B. TANG, *Depression risk prediction for chinese microblogs via deep-learning methods: Content analysis*, JMIR Medical Informatics, 8 (2020), p. e17958.

[197] X. WANG, C. ZHANG, Y. JI, L. SUN, L. WU, AND Z. BAO, *A depression detection model based on sentiment analysis in micro-blog social network*, in PAKDD, Springer, 2013, pp. 201–213.

[198] WHO, *World health organization.*, 2020.
[Online; accessed 2020-06-15].

[199] G. WHO, *Preventing suicide: a global imperative, 2014*, 2014.

[200] B. P. WILSON, *The phenomenon of grade inflation in higher education*, in Phi Kappa Phi Forum, vol. 79, National Forum: Phi Kappa Phi Journal, 1999, p. 38.

[201] J. WOLOHAN, *Estimating the effect of covid-19 on mental health: Linguistic indicators of depression during a global pandemic*, (2020).

[202] A. WONGKOBLAP, M. A. VADILLO, AND V. CURCIN, *Modeling depression symptoms from social network data through multiple instance learning*, AMIA Summits on Translational Science Proceedings, 2019 (2019), p. 44.

[203] WORLD HEALTH ORGANIZATION", *Mental health and covid-19*.

[204] WORLD HEALTH ORGANIZATION, *Coronavirus disease (covid-19) pandemic*, 2020.
[Online; accessed 2020-05-15].

[205] C.-S. WU, C.-J. KUO, C.-H. SU, S.-H. WANG, AND H.-J. DAI, *Using text mining to extract depressive symptoms and to validate the diagnosis of major depressive disorder from electronic health records*, Journal of affective disorders, 260 (2020), pp. 617–623.

[206] D. XEZONAKI, G. PARASKEVOPOULOS, A. POTAMIANOS, AND S. NARAYANAN, *Affective conditioning on hierarchical networks applied to depression detection from transcribed clinical interviews*, arXiv preprint arXiv:2006.08336, (2020).

[207] S. YANG, J. JIANG, A. PAL, K. YU, F. CHEN, AND S. YU, *Analysis and insights for myths circulating on twitter during the covid-19 pandemic*, IEEE OJCS, 1 (2020), pp. 209–219.

[208] W. YANG AND L. MU, *Gis analysis of depression among twitter users*, Applied Geography, 60 (2015), pp. 217–223.

[209] Z. YANG, Z. DAI, Y. YANG, J. CARBONELL, R. R. SALAKHUTDINOV, AND Q. V. LE, *Xlnet: Generalized autoregressive pretraining for language understanding*, in NIPS, vol. 32, 2019, pp. 5753–5763.

[210] Z. YANG, D. YANG, C. DYER, X. HE, A. SMOLA, AND E. HOVY, *Hierarchical attention networks for document classification*, in Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, 2016, pp. 1480–1489.

[211] A. YATES, A. COHAN, AND N. GOHARIAN, *Depression and self-harm risk assessment in online forums*, arXiv preprint arXiv:1709.01848, (2017).

[212] A. H. YAZDAVAR, H. S. AL-OLIMAT, T. BANERJEE, K. THIRUNARAYAN, AND A. P. SHETH, *Analyzing clinical depressive symptoms in twitter*, (2016).

[213] A. H. YAZDAVAR, H. S. AL-OLIMAT, M. EBRAHIMI, G. BAJAJ, T. BANERJEE, K. THIRUNARAYAN, J. PATHAK, AND A. SHETH, *Semi-supervised approach to monitoring clinical depressive symptoms in social media*, in ASONAM, ACM, 2017, pp. 1191–1198.

[214] H. YIN, S. YANG, AND J. LI, *Detecting topic and sentiment dynamics due to covid-19 pandemic using social media*, arXiv, (2020).

[215] J. YIN, Q. LI, S. LIU, Z. WU, AND G. XU, *Leveraging multi-level dependency of relational sequences for social spammer detection*, CoRR, abs/2009.06231 (2020).

[216] Y. YU AND X. WANG, *World cup 2014 in the twitter world: A big data analysis of sentiments in us sports fans,Äô tweets*, Computers in Human Behavior, 48 (2015), pp. 392–400.

[217] A. ZAFAR AND S. CHITNIS, *Survey of depression detection using social networking sites via data mining*, in 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, 2020, pp. 88–93.

[218] C. ZHANG, L. YANG, S. LIU, S. MA, Y. WANG, Z. CAI, H. DU, R. LI, L. KANG, M. SU, ET AL., *Survey of insomnia and related social psychological factors among medical staff involved in the 2019 novel coronavirus disease outbreak*, Frontiers in psychiatry, 11 (2020), p. 306.

[219] Y. ZHANG, N. ZINCIR-HEYWOOD, AND E. MILIOS, *Narrative text classification for automatic key phrase extraction in web document corpora*, in Proceedings of the 7th annual ACM international workshop on Web information and data management, 2005, pp. 51–58.

[220] J. ZHAO, L. GOU, F. WANG, AND M. ZHOU, *Pearl: An interactive visual analytic tool for understanding personal emotion style derived from social media*, in 2014 IEEE Conference on Visual Analytics Science and Technology (VAST), IEEE, 2014, pp. 203–212.

[221] W. ZHENG, L. YAN, C. GOU, AND F.-Y. WANG, *Graph attention model embedded with multi-modal knowledge for depression detection*, in 2020 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2020, pp. 1–6.

[222] J. ZHOU, S. YANG, C. XIAO, AND F. CHEN, *Examination of community sentiment dynamics due to covid-19 pandemic: a case study from australia*, arXiv:2006.12185 [cs], (2020).

[223] J. ZHOU, H. ZOGAN, S. YANG, S. JAMEEL, G. XU, AND F. CHEN, *Detecting community depression dynamics due to covid-19 pandemic in australia*, IEEE Transactions on Computational Social Systems, 8 (2021), pp. 982–991.

[224] H. ZOGAN, I. RAZZAK, S. JAMEEL, AND G. XU, *Depressionnet: Learning multi-modalities with user post summarization for depression detection on social media*, in Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 133–142.

[225] ——, *Depressionnet: Learning multi-modalities with user post summarization for depression detection on social media*, Proceedings of the 44rd International ACM SIGIR Conference on Research and Development in Information Retrieval, (2021).

[226] H. ZOGAN, I. RAZZAK, X. WANG, S. JAMEEL, AND G. XU, *Explainable depression detection with multi-modalities using a hybrid deep learning model on social media*, arXiv preprint arXiv:2007.02847, 25 (2020), pp. 281–304.

[227] ——, *Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media*, World Wide Web, (2022), pp. 1–24.

[228] H. ZOGAN, X. WANG, S. JAMEEL, AND G. XU, *Depression detection with multi-modalities using a hybrid deep learning model on social media*, CoRR, abs/2007.02847 (2020).

[229] M. L. ZOU, M. X. LI, AND V. CHO, *Depression and disclosure behavior via social media: A study of university students in china*, Heliyon, 6 (2020), p. e03368.