# Multiple instance learning for detecting depression markers in social media content

Wongkoblap, Akkapon

*Awarding institution:*
King's College London

# Multiple Instance Learning for Detecting Depression Markers in Social Media Content

Akkapon Wongkoblap

A thesis submitted in partial fulfilment for the

degree of Doctor of Philosophy

Department of Informatics

Faculty of Natural & Mathematical Sciences

King's College London

Dec 2020

# Abstract

Mental health problems are widely recognized as a major public health challenge worldwide. This highlights the need for effective tools for detecting mental health disorders in the population. Social media data is a promising source of information where people publish rich personal information that can be mined to extract valuable psychological cues. However, social media data poses its own set of challenges, such as the need to disambiguate between statements about oneself and about third parties. Traditionally, social media natural language processing (NLP) techniques have looked at text classifiers and user classification models separately. This presents a challenge for researchers who want not only to combine text sentiment and user sentiment analysis but also extract user's stories from textual content. The aim of this thesis is to develop a predictive model capable of detecting users with depression from their social media posts and identify textual content associated with self-disclosure of their mental health disorders on several platforms. The model needs to address the problem of anaphoric resolution and highlight anaphoric interpretations to focus on self-interpretation. This study proposes a multiple instance learning algorithm with (MILA-SocNet) and without anaphoric resolution (MIL-SocNet) to address the above challenges. The proposed algorithms were trained and tested on a microblogging platform, *Twitter*, and a social network platform, *Facebook* datasets. Several previously published models, ranging from classical machine learning to deep learning techniques, were applied to the same datasets and compared with our models. In our experiments on the microblogging and the social network dataset, our models outperformed alternative predictive models. Our proposed model with anaphoric resolution yields promising results relative to other predictive models and provides valuable insights into textual content relevant to the poster rather than a third party. Although the current performance of predictive models is in the ascendant, reliable predictive models will eventually allow early detection and pave the way for health interventions in the forms of offering relevant health services or delivering useful health information links. Finally, researchers, health organisations, social media providers, and governments, including social media users themselves, must not only prioritise individual and societal benefits but also eliminate risks to privacy to put forward the success of the profiling social media for mental health and digital interventions.

**Key words**: *Predictive Modelling, Machine Learning, Mental Health, Depression, Anaphora Resolution*

# Acknowledgements

I am grateful to everyone helping and supporting me throughout my PhD study. First of all, I would like to express my sincere appreciation and deepest thanks to my first supervisor, Dr. Vasa Curcin, who has continuously helped and provided me both academic and personal supports. His patience, motivation, and guidance are helpful and valuable to fill up my work.

My second thanks go to my second supervisor, Dr. Miguel A. Vadillo, for his suggestions concerning psychological and mental health knowledge. Without all of guidance and support from my primary and secondary supervisors, this thesis would not have been finished.

I am thankful to my colleague, Stefan Sarkadi, Francesca Mosca, Ionut Moraru, Xavier Ferrer Aran, Diego Sempreboni, Zhuoling Huang, Leonardo Cunha, Alex Burdusel, Parisa Zehtabi, and many others for joy, happiness, and assistance. Also, I appreciate all my Thai friends, who supported and accommodated, during my time in UK.

I would like to especially thank to Royal Thai Government and Suranaree University of Technology for a scholarship and an opportunity for pursuing my PhD.

Finally, thanks to my parents for their encouragement during my study and the writing of this thesis. This work would not have been completed without their inspiration, encouragement, and support over the years.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1
# Introduction

This thesis investigates the technology for predicting mental health problems from individuals' social network footprint. On the following pages, we will outline the motivation for the project and the impact of mental health disorders on the society, describe the objectives and contributions of this research, and, finally, provide a list of publications originating from this work.

## 1.1 Motivation

Mental health disorders are one of the most urgent health issues worldwide. According to statistics from the World Health Organisation (WHO), the number of global patients suffering from depression is estimated to be over 264 million in 2020 [1]. Mental illness impacts not only on the sufferers and their family and friends, but also causes a significant economic burden. The estimated cost of mental health services and treatments together with lost productivity at work in England was estimated at approximately 105 billion GBP in 2018 alone [2]. Globally, the costs of mental health problems were estimated at 2.5 trillion USD in 2010, and are expected to reach 6.1 trillion USD by 2030 [3]. A significant contributor to this cost is that people living with mental health problems sometimes receive inaccurate assessment [4]. Additionally, many mental health disorders can lead to suicidal ideation and suicide attempts [5].

Despite the large number of people suffering from mental health conditions, there is still a notable difficulty in gaining access to treatments and services, with only 7% and 28% of patients being able to access treatment services in developing countries and in developed countries, respectively [6, 7]. In low- and middle-income countries, between 76% and 85% of people with mental disorders receive no treatment [8].

These figures show that mental health problems are resonating across the society and demand new prevention and intervention strategies. Early detection of mental illness is an essential step in applying these strategies, with the diagnosis typically performed using validated questionnaires designed to detect patients with specific patterns of feelings, or social interaction [9–11]. Early detection can help plan strategic frameworks for public health and provide better direction for policies to improve nation’s health, e.g., providing more mental health services in the area of the large number of sufferers.

The rapid pace of technological advances enables individuals to easily and quickly access the internet, with the number of internet users reaching 4.57 billion in 2020 [12]. This increase in the number of people active on the internet has led to new types of social interactions via social media platforms. This research is focused on two subsets of these platforms: *social networking sites* (e.g., Facebook), and *microblogging sites* (e.g., Twitter) which tend to attract the largest number of active users - *Facebook* has over 1.88 billion daily active users, and there are 199 million daily active accounts on *Twitter* at the time of writing this thesis [13, 14].

Unlike other types of social media, social network and microblogging platforms, so-called *profile-based social media*, allow their users to create personal profiles, establish new relationships as well as maintain close friendships, publish status updates, and interact with other users. Profile-based social media users openly express a variety of thoughts, feelings, and emotions every day, creating a vast volume of data through those activities.

Social media data enables the use of data mining and machine learning algorithms to detect meaningful patterns and gain new insights. Data science enables researchers to interpret and visualise information from complex datasets. User-generated content on social media, review, blog, and messaging board platforms offers an opportunity for researchers to explore and classify a large number of content not only in health [15–17], but also in domains such as marketing [18] and politics [19].

Within the health domain, profile-based social media platforms are starting to attract researchers' attention as a means to obtain new knowledge about the symptoms and predictors of health and mental health problems, by analysing the feelings, thoughts, and activities described in users' posts. With some users publicly talking about their mental health on their social network profiles, it becomes possible to train classification engines to detect online users with mental health problems [20, 21]. Using microblogging data, in particular, studies have looked into users with depression [22–25], postpartum depression [26], anxiety, obsessive compulsive disorder (OCD), and post-traumatic stress disorder (PTSD) [22, 27]. Social network data were used to detect users with depression [28, 29] and postpartum depression [30].

Generally, text classifiers and user classification models tend to be developed separately. This presents a challenge for researchers who want to understand both text sentiment and user sentiment analysis simultaneously. In this thesis, we present a predictive model capable of detecting users with depression and instantly identifying their messages as health related. An ideal technique to develop this kind of model is multiple instance learning (MIL) [31], where the model can learn from only a set of labelled bags/users instead of a set of individual instances/user-generated messages.

A challenge posed by social media is that posts may frequently refer to individuals other than the users themselves [32]. Anaphora resolution is an established natural language processing (NLP) problem and an emerging field in the analysis of social media content that helps determine which previously mentioned person is the subject of a subsequent statement and understand references to someone in content on social media.

To the best of our knowledge, no previous study has focused on *the seamless integration of MIL and anaphora resolution*, in the field of profiling social media users for mental health. MIL can be used to develop a predictive model for detecting users suffering from mental disorders on profile-based social media and identifying textual content related to self-disclosure of mental health. With the help of anaphora resolution, a predictive model can identify user-generated content, statements, thoughts and attitudes relating to mental health self-disclosure of the user rather than a third party. Our proposed method may give us a resource for better understanding symptoms of mental disorders.

## 1.2 Aim and Objectives

The overall aim of this research is to investigate the feasibility of using machine learning techniques to detect users suffering from depression using their profile-based social media content and identify textual content associated with self-disclosure of their mental health disorders. To address this aim, we set the following objectives:

1. Survey the scope and limits of cutting-edge techniques for developing predictive models to identify profile-based social media users with mental health disorders.
2. Construct predictive models to classify profile-based social media users with mental health illnesses and identify textual content related to self-disclosure of mental illness.
3. Evaluate the feasibility of these models on examples of a microblogging platform and a social network platform.
4. Examine the impact of the work presented on ethical issues concerning the use of social media data for research.

## 1.3 Contributions

On the basis of these objectives, this thesis makes the following contributions:

- **A systematic review**: A systematic literature review is conducted using keywords to search articles published on data mining and machine learning in the context of common mental health disorders from profile-based social media data. The review provides information about the scope and limits of cutting-edge techniques for predictive analytics in mental health and about research gaps, in this area of research. This is presented in *Chapter 2*.

- **A set of effective predictive models**: This research develops a set of predictive models for classifying users with mental health disorders from profile-based social media data. The models can also detect the users and instantly identify posts revealing self-disclosure of mental health disorders. *Chapter 3* presents the architecture of the classifiers. The models are evaluated based on two different datasets: one from a microblogging platform explained in *Section 4.1* and another from a social network platform described in *Section 4.2*. At the end of *Chapter 4*, we compare the results from our proposed models against a set of previously published models and highlight important discussion.

- **A tool for gathering data from a microblogging and a social network platform**: This is a sub-contribution from our study presented in *Appendix A*. This type of research tasks requires a new data collection approach to collect profile-based social media data from participants. The thesis follows other studies in this field to develop a tailored tool compatible with a social media platform, using the Application Programming Interfaces (APIs) provided by a social media platform to automatically pool user's data - with their permissions, a permanently accessible opt-out option.

- **An ethical framework of profiling profile-based social media for mental health**: *Chapter 5* surveys ethical concerns and issues of the use of profile-based social media data for mental health analytics. It provides guidelines on how to conduct research on this area and explains solutions to avoid ethical issues.

- **A framework for a personalised intervention model**: This is a sub-contribution from our study presented in *Chapter 6*. A framework for personalised intervention promoting mental health services, providing online help, or delivering useful health information links is developed to foster future research. The personalised intervention model is expected to capture users' data and provide valuable help to target users with mental illness.

## 1.4 Publications

- A. Wongkoblap, M. A. Vadillo and V. Curcin, "Researching mental health disorders in the era of social media: Systematic review," *J Med Internet Res*, 19(6): e228, Jun 2017.

  The systematic review explored the scope and limits of the state-of-the-art in predictive analytics in mental health and to review associated issues, such as ethical concerns. This paper is reported in Chapter 2.

- A. Wongkoblap, M. A. Vadillo and V. Curcin, "Modeling Depression Symptoms from Social Network Data through Multiple Instance Learning," *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science,* San Francisco, CA, 2019, pp. 44-53.

  This paper proposed a multiple instance learning model for detecting depressed users from a social network platform. The dataset used in this study was taken from myPersonality project. This publication is related to Chapter 4.

- A. Wongkoblap, M. A. Vadillo and V. Curcin, "Predicting Social Network Users with Depression from Simulated Temporal Data," *IEEE EUROCON*

*2019 -18th International Conference on Smart Technologies*, Novi Sad, Serbia, 2019, pp. 1-6.

This paper developed a multiple instance learning model for classifying users with depression from a social network platform.

- E. Ford, K Curlewis, A. Wongkoblap, and V. Curcin, "Public Opinions on Using Social Media Content to Identify Users with Depression and Target Mental Health Care Advertising: Mixed Methods Survey," *JMIR Mental Health 2019*. 6(11): e12942.

  Akkapon Wongkoblap was a co-author in this article. This paper surveyed public opinions and conducted a group interview on profiling profile-based social media for mental disorders. This is mentioned in Chapter 5.

- A. Wongkoblap, M. A. Vadillo and V. Curcin, "Classifying Depressed Users with Multiple Instance Learning from Social Network Data," *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, New York, NY, 2018, pp. 436-436.

  This produced a preliminary study on implementing a multiple instance learning model for predicting depressed users from the myPersonality dataset collected from a social network platform. Chapter 4 is related to this paper.

- A. Wongkoblap, M. A. Vadillo and V. Curcin, "Depression Detection of Twitter Posters using Deep Learning with Anaphora Resolution," *JMIR Mental Health*.

This paper developed a multiple instance learning model with anaphora resolution for classifying users with depression from a microblogging platform. This is the core content in Chapter 4.

- A. Wongkoblap, M. A. Vadillo and V. Curcin, “Social Media Big Data Analysis for Mental Health Research,” in *Mental Health in the Digital World*. *[In press]*

  This book chapter covers the use of data from social media platforms such as microblogs, e.g., *Twitter*, and social networks, e.g., *Facebook*, to advance mental health research and practice. Content of this paper is mainly presented in Appendix B.

## 1.5 Ethics Approval

This study reported objectives and methodology to College Research Ethics Committee. This study was approved by King's College research ethics committee (reference number LRS-16/17-4705).

## 1.6 Thesis Structure

The thesis is structured as follows:

- **Chapter 2:** Provides the systematic review that explored the scope and limits of the state-of-the-art in mental health predictive analytics and associated issues

- **Chapter 3:** Describes the architecture of our predictive models for classifying users with depression from their profile-based social media data
- **Chapter 4:** Evaluates the proposed machine learning models based on two different datasets: one from a microblogging platform and another from a social network platform
- **Chapter 5:** Surveys ethical concerns and issues of the use of profile-based social media data for mental health analytics, and provides guidelines to avoid the ethical issues
- **Chapter 6:** Discusses the main contributions of this thesis and presents future directions.
- **Appendix A:** Introduces the necessary background for mental health disorders, and the development of predictive models from social media platforms
- **Appendix B:** Presents the data capture tools developed to collect data from profile-based social media platforms: *social networking sites* (e.g., Facebook), and *microblogging sites* (e.g., Twitter)

# Chapter 2
# Literature Review

This chapter presents the state-of-the-art in development of predictive models for detecting users with mental illness from profile-based social media data, covering the last ten years of research since the first efforts in the field. A systematic literature review was performed, using keywords to search articles on data mining in profiling mental health from social media. The chapter begins with the summary of our systematic review and then presents detailed information about predictive model construction and identifies gaps in predictive analytics in mental health for profile-based social media data.

## 2.1 Systematic Review

A systematic literature review was performed in September 2020 using keywords to search articles focused on the prediction of mental health problems based on data from profile-based social media between 2010 and September 23, 2020 in medical and computer science databases including *PubMed*, *Institute of Electrical and Electronics Engineers (IEEE Xplore)*, *Association for Computing Machinery (ACM Digital Library)*, *Web of Science*, and *Scopus*.

The titles and abstracts of articles were filtered using search terms. The initial search returned a total of 10,562 papers. Only articles published in peer-reviewed journals and written in English were included. Further inclusion criteria were that studies had to (1) focus on predicting mental health problems through profile-based social media data and (2) investigate prediction or classification models based on users' text posts, network interactions, or other features of profile-based social media platforms. This review focused on profile-based social media platforms—that is, those allowing users to create personal profiles, post content, and establish new or maintain existing relationships. This means that *Reddit* does not fit into our study. This is because it allows users to update social news, rate web content, and make discussion among users, as explained in *the appendix A.2.1*.

Studies were excluded if they (1) only analysed the correlation between profile-based social media data and symptoms of mental illness, (2) analysed textual contents only by human coding or manual annotation, (3) examined data from online communities (e.g., LiveJournal), (4) focused on the relationship between social media use and mental health disorders (e.g., so-called Internet addiction), (5) examined the influence of cyberbullying on mental health, or (6) did not explain where the datasets came from.

Following a careful analysis of the titles, abstracts and content of papers, 82 papers were selected for our review. In addition, the reference lists of included articles were examined for additional sources. Furthermore, the proceedings in several conferences and journals were manually searched to find additional articles that the search terms might have excluded. Figure 2-1 depicts the preferred reporting items for systematic reviews and meta-analyses (PRISMA) flow diagram of the results of searching and screening articles following the above search methodology.

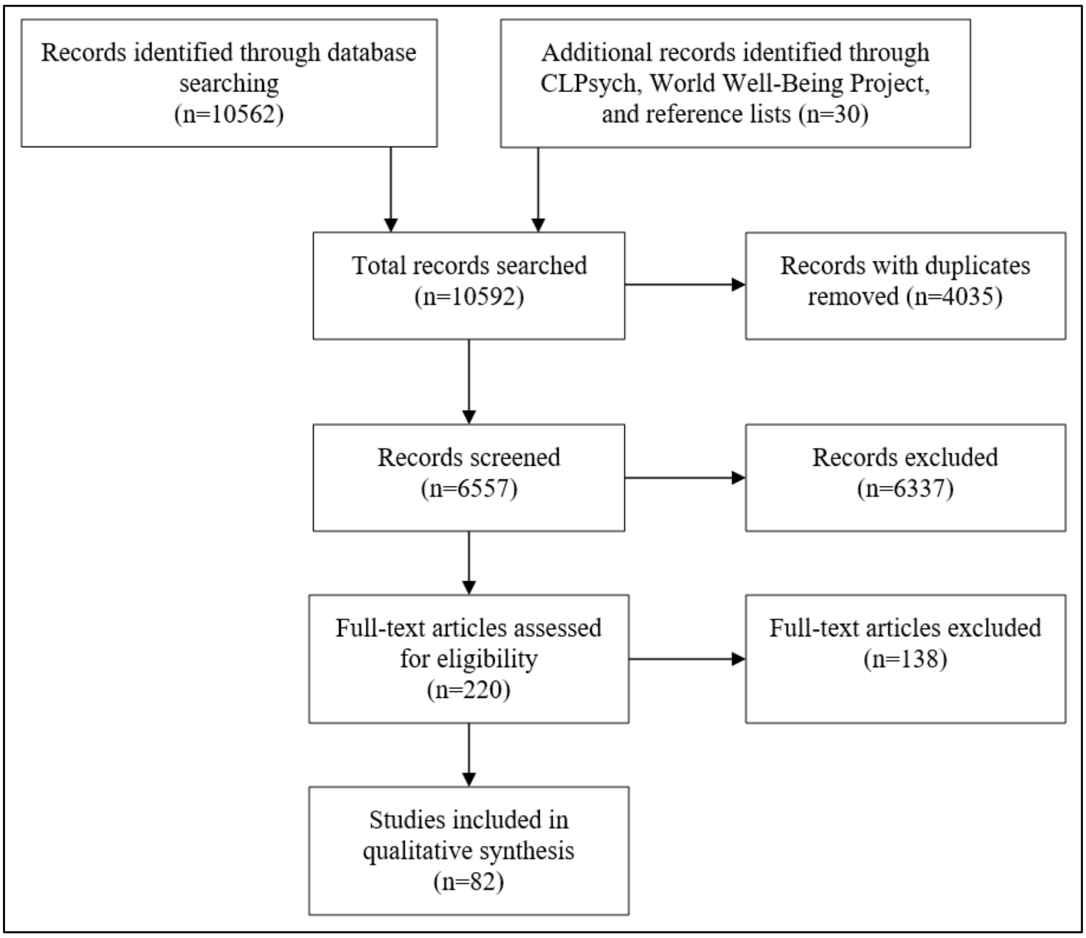

Records identified through database searching (n=10562)
Additional records identified through CLPsych, World Well-Being Project, and reference lists (n=30)
Total records searched (n=10592)
Records with duplicates removed (n=4035)
Records screened (n=6557)
Records excluded (n=6337)
Full-text articles assessed for eligibility (n=220)
Full-text articles excluded (n=138)
Studies included in qualitative synthesis (n=82)

*Figure 2-1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram. CLPsych: Computational Linguistics and Clinical Psychology Workshops.*

## 2.2 Mental Health Disorders Studies from Social Network Data

The selected studies can be divided into several distinct categories. Several studies [21, 22, 37–46, 24, 47–56, 27, 57–66, 28, 67–74, 29, 33–36] used datasets from social networks to examine depression. Postpartum depression disorder was explored by De Choudhury et al. [26, 30]. PTSD was studied by [22, 27, 74, 75, 37, 39–42, 61, 63, 65]. Anxiety and OCD were investigated by [27, 65, 76]. Borderline disorder and bipolar disorder were studied by [22, 27, 74, 77]. Seasonal affective disorder (SAD) were studied by Coppersmith et al. [22, 27] and Chen et al. [74]. Eating disorder were explored by Chancellor et al. [78], Coppersmith et al. [22], Wang et al. [79], and Prieto et al. [44]. Attention Deficit Hyperactivity Disorder (ADHD), anxiety, and schizophrenia were examined by Coppersmith et al. [27], panic disorder were investigated by [65], and sleep disorder were studied by Jamison-Powell et al. [80]. Users with suicidal ideation were studied by [64, 81, 90–95, 82–89]. Happiness, satisfaction with life and well-being were studied by [96–104].

## 2.3 Studied Language

Of the studies included in this review, 52 papers analysed profile-based social media contents written in English [21, 22, 38–43, 50, 51, 54, 58, 24, 59–63, 65, 66, 68, 72, 73, 26, 74, 75, 77–82, 84, 85, 27, 88, 89, 91–93, 96, 97, 101, 102, 104, 28, 105, 29, 30, 33, 37]; 17 studies investigated Chinese text [35, 45, 83, 86, 87, 90, 95, 99, 100, 47, 48, 56, 64, 67, 69, 71, 76]; 3 studies explored Russian content [52, 53, 57]; 2 experiments focused on Korean [34, 46], 2 on Japanese text [36, 49], 1 on Arabic content [55]; 1 looked at Turkish content [98]; 1 at Italian text [103]; 1 at Spanish [94]; 1 at Portuguese [70]; and one jointly at Spanish and Portuguese [44].

## 2.4 Processes of Predictive Modelling

The 82 papers selected were categorised and summarised the processes of predictive modelling, which can be divided into six steps. The first step is data collection, being considered the most important step, because data is essential to train predictive models. Data are cleaned and preprocessed to ensure that they are in the form required by the analytical algorithms. Then, key features related to the research domain are prepared for model construction. Overall, this involves *data collection*, *data preprocessing*, *feature extraction* and *feature selection*, *training a set of predictive models* and *validating predictive models* (see Figure 2-2).

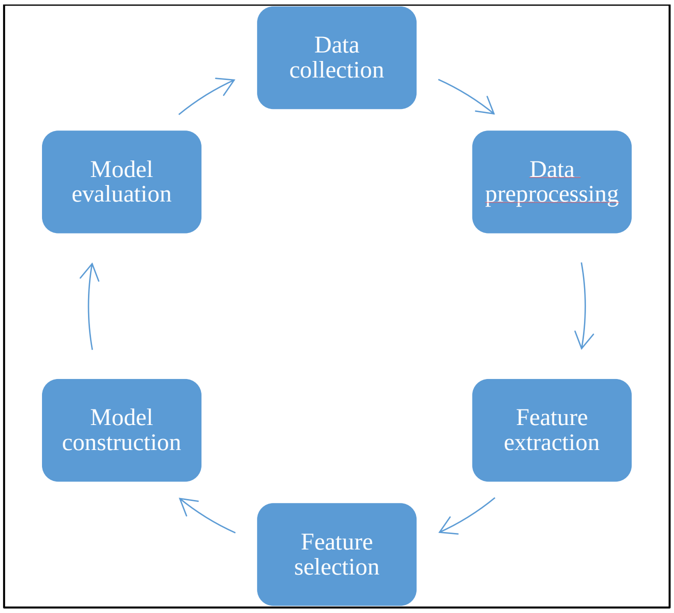


*Figure 2-2. The processes of Predictive modelling*

### 2.4.1 Data Collection Techniques

In the context of the present project, datasets are directly or indirectly obtained from profile-based social media. There are two broad approaches to data collection: (1) collecting data directly from the subjects with their consent using surveys and electronic data collection instruments (e.g. Facebook apps), and (2) aggregating data extracted from public posts.

In previous studies predicting mental health problems from profile-based social media data, the specific methods of collecting data directly from participants varied with the purpose of each study and the target platform. Experimental designs for the studies included posting project information on relevant websites inviting participants to take part in the project [30, 36, 64, 70, 83, 103], and posting tasks on crowdsourcing platforms asking for project volunteers [21, 24, 50, 60, 63, 81, 96]. In the latter case researchers would post detailed tasks and study requirements on platforms such as Mechanical Turk [106] to attract subjects. As part of a questionnaire, the participants would typically be asked to provide informed consent allowing collection of their social media data.

A range of questionnaires were used to collect data on participants' levels of depression and life satisfaction, including CES-D [21, 24, 34, 36, 46, 50, 51, 55, 63, 67], the Patient Health Questionnaire-9 (PHQ-9) [30, 55], Beck's Depression Inventory (BDI) [24, 34, 36, 46, 52, 53, 57, 60, 70], Zung Self-Rating Depression Scale (Zung's SDS) [49], the Depressive Symptom Inventory–Suicide Subscale (DSI-SS) [81], and Symptom Checklist-90-Revised (SCL-90-R) [76]. The instruments used to detect suicidal ideation and possibility of individuals' suicide were the Suicidal Possibility Scale (SPS) [64, 83, 86, 95], the Acquired Capability for Suicide Scale (ACSS)

[81], Depression Anxiety Stress Scales-21(DASS-21) [64], the Trauma Screening Questionnaire (TSQ) [63], and the Interpersonal Needs Questionnaire (INQ) [81]. Satisfaction with life and well-being were measured with the Satisfaction with Life Scale (SWLS) [96, 97], the Positive and Negative Affect Schedule (PANAS) [100], and the Psychological Well-Being Scale (PWBS) [100]. One study used the Revised NEO Personality Inventory (NEO-PI-R) to assess personality [29, 51, 60]. One study used medical record from a hospital to classify users [28].

The second approach is to pool only public posts from profile-based social media platforms, by using regular expression to search for relevant posts, e.g. "I was diagnosed with [condition name]" [22, 27, 62, 65, 66, 68, 71–74, 89, 92, 38, 93, 94, 39, 54–56, 58, 59, 61]. A study collected profiles of users committing suicide from *Hereafter*, which contains last posts of the users [91].

To collect social network data, each data source required a custom capture mechanism, due to a lack of standards for data collection. Facebook-based experiments gathered user datasets by developing custom tools or web apps connecting to the Facebook APIs [28, 30, 34, 46, 67, 103]. Another group of studies used Twitter APIs to explore cues for mental disorders [22, 24, 50, 54, 55, 58–61, 63, 66, 72, 26, 73, 75, 77, 79–82, 84, 85, 88, 29, 89, 91–95, 101, 102, 104, 33, 36, 43–45, 49]. A similar approach was adopted for Instagram APIs [21, 56, 68, 70, 78], Sina Wiebo APIs [35, 45, 90, 99, 100, 47, 48, 64, 69, 71, 83, 86, 87], and Vkontakte APIs [52, 53, 57].

Another way of obtaining data was promoted by the *myPersonality* project, which provides both social network data and a variety of psychometric tests' scores for academic researchers [107], and was used by 5 studies [29, 51, 60, 96, 97]. The myPersonality project was shut down in 2018[1]. Some studies originated from workshops where the organisers shared data already approved by an Institutional Review Board (IRB) for analysis [37, 40–42].

### 2.4.2 Data Preprocessing

The corpus of data is typically preprocessed by (a) removing unsuitable samples and (b) cleaning and preparing the data for analysis. Information and questionnaires from participants might contain useless data and incomplete details, which are typically removed from studies in order to improve the accuracy of prediction and classification results. In previous studies on this topic, participants who took an abnormally short or long time to complete the questionnaires were excluded from studies [24, 36, 50, 86]. Low-activity participants who had published less than a defined threshold of posts were removed from studies [22, 28, 77, 83, 86, 97, 100, 102, 103, 35, 52, 53, 56, 59–61, 63]. Participants with poor correlations among two different questionnaires were excluded from the final dataset in studies [24, 36].

As part of the data cleaning process, each post was checked for the majority written language (e.g., contained at least 70% English words [22, 27, 38, 59, 75, 80, 96]). This ensures that the available tools are suitable to analyse the posts. Each post was preprocessed by eliminating stop words and irrelevant data (i.e., retweets, hashtags, URL), lowercasing characters, and segmenting sentences [40, 42, 62, 67, 68, 71, 72, 75, 82, 88, 91, 94, 44, 50, 52–55, 58, 61]. Emoticons were converted to other forms such as ASCII codes [41, 58, 68, 103] to ensure data was machine-readable.

[1] https://sites.google.com/michalkosinski.com/mypersonality

Anonymization was also performed to remove any potentially identifiable usernames [21, 43, 63, 75, 80, 82, 84, 85].

### 2.4.3 Feature Extraction

There are many potential techniques to extract features that could be used for predicting mental health problems in profile-based social media users. Several studies have attempted to investigate the textual contents of profile-based social media platforms to understand what factors contain cues for mental disorders. However, some research projects have used alternative methods. In general, previous studies relied on three broad approaches to feature extraction: text analysis, image analysis, and social interaction.

In text mining, sentiment analysis is a popular tool for understanding emotion expression. It is employed to classify the polarity of a given text into categories such as positive, negative and neutral [108]. Several studies [24, 26, 60, 63, 64, 66, 74, 75, 79–81, 83, 28, 86, 88, 91, 94, 96, 97, 99–101, 103, 30, 105, 43, 45, 50, 51, 54, 59] used the well-known Linguistic Inquiry and Word Count (LIWC) to extract potential signs of mental problems from textual content (as explained in *the appendix A.4.3*).

There are alternative tools used to extract features. *OpinionFinder* [109] performs subjectivity, objectivity, and sentiment analysis, which was used by Bollen et al. [102]. *SentiStrength* [110] assesses the polarity between positive and negative words and the levels of strength of positive and negative words in a textual message, was used by Kang et al. [33] and by Durahim and Coşkun [98]. Word embeddings was used by [54, 56, 73, 90, 93, 94, 58, 61, 62, 67–71].

Custom tools were also developed for performing sentiment analysis. *Latent Dirichlet allocation (LDA)* [111] is a useful and powerful technique to create topic models. LDA analyses latent topics, based on word distribution, and then assigns a topic to each document. *Topic modelling* was employed in studies [28, 36, 54, 60, 78, 86, 92, 96] to extract topics from user-generated posts. *Affective Norms for English Words (ANEW)* [112] was used to qualify the emotional intensity of English words in [26, 50, 63].

Social media posts tend to be rich in emoticons. As a consequence, several studies [33, 47, 55, 68, 91] have looked into the meaning and mood states associated with their use.

Apart from posting text messages, profile-based social media platforms allow users to post images. Some studies investigated these images for cues of mental disorders [21, 33, 70, 90, 94, 45, 54, 56–58, 60, 66, 68]. Colour compositions and SIFT descriptors techniques were utilised to extract the emotional meaning of each individual image [33]. Image properties, including colour theme, saturation, brightness, colour temperature, and clear or dull colour were analysed in [21, 45, 54, 57, 66]. Capturing emotions from images was used by [66].

Finally, profile-based social media platforms contain millions of interaction and relationships among users. Profile-based social media users not only can connect and add online friends, but also can post, comment, and reply to their friends. The resulting graph structure, comprising information about interactions, relationships, and friendships, was mined to understand the cues that can be connected to symptoms of mental disorders (e.g., interactions among depressed users and assortative mixing patterns) [48, 51, 66, 79, 102].

### 2.4.4 Feature Selection

Feature selection isolates a relevant subset of features that are able to predict symptoms of mental disorders or correctly label participants, while avoiding overfitting. Statistical analysis is typically performed to discover a set of parameters that can differentiate between users with mental disorders and users without mental disorders. The techniques used in the selected studies were Pearson's correlation coefficient [34, 44, 100], ANOVA [66], Correlation-based feature selection (CFS) [44], Spearman rank correlation coefficient [46], and Mann–Whitney U test [46, 94]. The dimensionality of features was reduced by Principal Component Analysis (PCA) in studies [24, 26, 29, 50, 52, 57, 85], Randomised Principle Component Analysis (RPCS) in [96], Forward Greedy Stepwise in [35], Binary logistic regression in [47], Gain Ratio in [44], and Relief techniques in [44].

### 2.4.5 Predictive Model Construction

In the selected studies, prediction models were used to detect and classify users according to mental disorders and satisfaction with life. To build a predictive model, a selected set of features is used as training data for machine learning algorithms to learn patterns from those data.

A common approach to build predictive models in profiling social media users with mental health problems is supervised learning, where the sample data contains both the inputs and the labelled outputs (as mentioned in *appendix A.5.1.1*). The model learns from these to predict unlabelled inputs from other sources and provide prediction outputs [113].

The techniques used in previous studies include Support Vector Machine (SVM) [36, 38, 74, 76, 83–85, 44, 51–53, 57, 59, 64, 71], Linear SVM [33, 37, 42, 51, 55, 72, 79, 88], SVM with a Radial Basis Function (RBF) kernel [24, 26, 33, 42, 50, 72, 79, 87]. Regression techniques included Ridge regression [96], Linear regression [29, 35, 60], Loglinear regression [22, 74, 75], Logistic regression [28, 35, 95, 104, 105, 51, 52, 57, 61, 71, 82, 84, 87], Binary logistic regression with Elastic Net regularisation [37, 39], Linear regression with stepwise selection [49, 86, 100], Stepwise logistic regression with forward selection [30], Regularised multinomial logistic regression [78], Linear Support Vector Regression (SVR) [41, 100], Least Absolute Shrinkage and Selection Operator (LASSO) [100, 101], and Multivariate Adaptive Regression Splines (MARS) [100]. Other algorithms used for binary classification covered Decision trees [44, 47, 48, 71, 72, 74, 85, 87, 91, 92], Random forest [21, 51, 77, 87, 89, 91, 95, 103, 52, 53, 55, 57, 59, 63, 71, 74], Rules decision [47], Naïve Bayes [44, 47, 85, 87, 91, 52, 55, 57, 71, 72, 74, 76, 79], Nearest Neighbour (kNN) [44, 51, 57, 79, 92], Maximum Entropy (MaxEnt) [38], Classification and Regression Trees [51], Ensemble learning [52, 54, 55, 57, 66, 91], Neural network [51, 52, 57, 65, 76], Deep Learning neural network (e.g., CNN and LSTM) [45, 56, 93, 94, 58, 62, 67–70, 73, 90], MIL [70], and Reinforce learning [73].

From the above list of specific model use of note, most popular models used in this field were SVM, logistic regression, and random forest for binary classification. Some studies used linear regression, LASSO, and SVR to predict continuous values such as the severity levels of mental, rather than classes (e.g., non-depressed and depressed classes). Relatively few studies used deep learning approaches to develop predictive model for classifying users with mental health problems based on their social media content.

### 2.4.6 Model Evaluation

Following the model construction, the model accuracy must be measured using a test dataset. The most common technique was n-fold cross-validation. Studies [24, 27, 55, 60, 61, 66, 67, 70, 72, 76, 77, 82, 28, 85–87, 91, 93, 94, 103, 105, 35–39, 44, 51] employed 10-fold cross validation to verify their prediction models and classifiers, whilst 5-fold cross validation was used by studies [45, 52, 89, 95, 100, 53, 54, 57, 62, 63, 65, 73, 79]. Leave-one-out cross validation was used in [22, 59, 74, 81].

With the use of the cross-validation method to generalise and measure the predictive performance of models, the above studies typically compared their novel and replicated models in terms of accuracy, precision, recall, and F1-score along with ROC curves to visually display their performances in a single graph. This helps authors of the papers to easily compare and find the best model in their studies.

The performance of predictive models can also be evaluated in other datasets. Studies [29, 33, 61–63, 78, 84, 88, 89, 96, 101, 103, 41, 105, 42, 49, 51, 52, 56, 57, 59] divided the collected dataset into training and test subsets to measure the accuracy of their models. Studies [50, 54, 62, 71, 73, 75, 95, 98, 99] collected a new dataset to evaluate the accuracy of the predicted results and compare the predicted results with a set of known statistics (e.g., the depression rates in US cities, student satisfaction survey, and Gross National Happiness percentages of provinces of Turkey).

## 2.5 Anaphora Resolution

Anaphora resolution is a well-known natural language processing problem and also an emerging field that helps to determine which previously mentioned person is the subject of a subsequent statement. Reference resolution in sentiment analysis has been developed in many fields including machine translation, paraphrase detection, summarisation, question answering and sentiment analysis. There are three reference resolution algorithms [114]:

1. Rule-based entity resolution is an NLP technique to extract syntactic rules and semantic knowledge from a text. This approach requires hand-crafted methods to extract those features.
2. Statistical and machine learning based entity resolution is a set of learning-based and probabilistic methods to understand the co-reference of a reference referring to an early entity.
3. Deep learning for entity resolution has been developed to reduce hand-crafted features requirements. It represents words as vectors conveying sematic units.

We searched for studies that applied anaphora resolution into profile-based social media data. We found that only one study from Aktaş et al. [32] investigated anaphora resolution for conversations in a microblogging platform by using a corpus and manual annotation. They found that conversations on the platform contained anaphora relations and it was possible to apply an anaphora resolution technique on the data.

## 2.6 Discussion

The purpose of this chapter was to investigate the state of the art of development of research on machine learning techniques predicting mental health from profile-based social media data and anaphora resolution for conversations on profile-based social media platforms. This review also focused on identifying gaps in research and potential applications to detect users with mental health problems.

### 2.6.1 Data Collection

Several sites were used as sources of data. Facebook is possibly the most popular social network platform. However, only a few studies relied on Facebook datasets to predict mental disorders. One reason for this might be that, by default, users on this site do not make their profiles publicly accessible. Another reason is that getting data from Facebook requires consent from users and review processes (as explained in *appendix A.3.1*).

Twitter has been a popular source of profile-based social media data in the surveyed articles, as it provides two ways of accessing the public data: retrospective (using their search APIs collecting historical data) and prospective (via their streaming APIs retrieving real-time data). With the new updated version 2 of the API released in 2020, designed to foster research especially in the field of social media analytics, researchers can now collect bigger datasets and more real-time data, which may contribute to wider adoption of these methods in mental health research.

In terms of data collection from users, there are some differences between obtaining data through participants' consent and using regular expression to search for relevant posts. The former option can provide us the real results of the prevalence of mental disorders from participants. The latter approach reduces the time and cost of identifying users with mental illness [22].

### 2.6.2 Feature Extraction Techniques

The LIWC tool is mostly used for text analysis in psychological research. It extracts many category features, such as style words, emotional words, and parts of speech, from textual contents. It is relatively easy to use and does not require programming skills. Users can just select and open a file or a set of files and LIWC will extract the relevant features and values of each feature. However, there are some disadvantages too. First, LIWC is a proprietary software and users have to purchase a license to use it. Second, the feature database of the tool is not easy to modify. To do this, researchers might need programming skills.

To overcome these shortcomings, there are alternative tools to extract features. However, these tools are rather limited in that they can extract only some features. WordNet is an English large lexicon that can be used to extract parts of speech from text and find semantic meaning of words [115]. Mallet is a useful NLP tool to classify or cluster documents, create topics, and perform sequence labelling [116]. Each word from assigned text can be tagged with parts of speech by POS Tagger [117]. With the advance of natural language processing, word embedding is one of the useful and successful techniques to deal with natural language process problems.

### 2.6.3 Changes in Research Methods over Time

Despite the thousands of articles collected through our search terms, the results of our review suggest that there is a relatively small but growing number of studies using machine learning models to predict mental health problems from profile-based social media data. From the initial set of matched articles, only 82 papers met our inclusion criteria and were selected for review (see Figure 2-3). Some of the excluded studies focused on analysis of the effects of profile-based social media use on mental health and well-being states of individual users, and the influence of cyberbullying in profile-based social media platforms on other users.

A wide range of machine learning algorithms were used in the reviewed studies. Relatively few studies used deep learning algorithms to build a classifier, with the rest relying on classical machine learning techniques such as SVMs, regression models, and decision trees to build classification models. This can be noticed from Figure 2-3. Papers published before 2017 intensively focused on classical machine learning techniques. With the rise in popularity and success of deep learning techniques, publications published after 2017 increasingly relied on deep learning techniques.

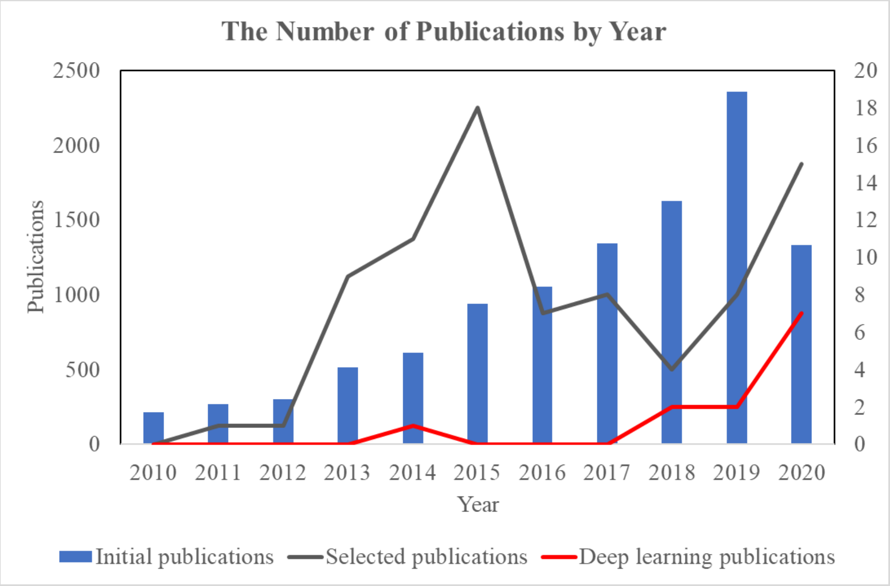


*Figure 2-3 The number of publications from 2010 to 2020. The y-axis on the left-hand side presents the number of initial papers from databases. The y-axis on the right-hand side shows the number of selected papers and publications using deep learning techniques.*

### 2.6.4 What Can Be Improved in This Research?

With the immense advancements and the success of machine learning and NLP, there is still some room for the improvement in this research area. The evolution of predictive models from classical machine learning to deep learning techniques provides promising tools to classify online users with mental health issues. However, deep learning is a black box, as opposed to human-interpretable models, such as regression and decision trees, raising the issue of whether it is possible, or indeed necessary, to have these algorithms validated by clinical experts [118].

Another method can be used to improve the predictive performance is an ensemble model, which combines multiple diverse base models. Each of the base models is used to focus on its specific outcome e.g., the first model for detecting pictures related to mental health and the second model for classifying text associated with a health topic. Even the ensemble model consists of varied base models, it is used to perform one final task e.g., classifying a user with depression from social media data [119].

Many recent approaches of NLP have not been applied in this research area. One successful example of NLP in this context is the bidirectional encoder representations from transformers (BERT), which is pre-trained on deep bidirectional representations from unlabelled textual content by conditionally connecting on both left and right context [120].

With the identified research gaps and the recommended improvement, the present research plans to develop a deep learning model with MIL that can provide information of how the model makes a decision to classify users with mental health disorders.

Another point to highlight is that no study has used anaphora resolution on profile-based social media posts to detect users with mental health disorders. This is an interesting and challenging task to *develop a predictive model that can detect users with mental disorders and identify textual content relevant to the self-disclosure and mental health topics*.

## 2.7 Summary

The purpose of this chapter was to provide an overview of the state-of-the-art in research on machine learning techniques for predicting users with mental health from profile-based social media platforms. This helps to identify gaps in the area and frame methodological solutions to fill the gaps. Most of the selected studies approached this problem using text analysis. Predictive models and binary classifiers can be trained based on features obtained from all techniques explained above. Moving forward, a deep learning model is a possible solution to improve performance of a predictive model, as it may help to extract insightful information of how the model thinks. The next chapter will explain our proposed predictive models.

# Chapter 3

# Multiple Instance Learning from Profile-Based Social Media Platforms

After reviewing studies on classifying users with mental health problems from social media data, we found that there is a need for predictive models which can provide insights into how the model makes a decision. This chapter describes the architecture of our predictive models for classifying depressed users from their profile-based social media data. We explain how MIL models with supervised neural networks classify depressed users and identify posts related to self-disclosure of mental disorders.

## 3.1 Notation and Operations

Before beginning this chapter, we will provide notation and operations used throughout the rest of the thesis. First, we will start with types of variables, and then describe mathematic operations.

### 3.1.1 Notation

There are two main types of variables used throughout this chapter. First, *a vector* is an array of numbers. We use a lowercase letter to represent a vector variable such as $x$. We can access each element of the vector by using its variable name with a subscript index such as $x_1, x_2$ or $x_n$. Elements of a vector are illustrated as a column or a row enclosed in square brackets:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}, x = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_n \end{bmatrix}$$

Second, *a matrix* is a 2-D array or a rectangular array of numbers notated by a uppercase letter such as $W$. To identify an element of a matrix, we use its matrix name with subscript indexes such as $W_{1,1}, W_{2,1}$ or $W_{m,n}$. To present elements of a matrix, we use rows and columns enclosed in square brackets:

$$W = \begin{bmatrix} W_{1,1} & \cdots & W_{1,n} \\ \vdots & \ddots & \vdots \\ W_{m,1} & \cdots & W_{m,n} \end{bmatrix}$$

### 3.1.2 Operations

This section describes important operations used throughout this chapter. We will start with a simple operation, which is *vector concatenation*. This thesis denotes the operation of concatenating vectors using vector names in square brackets such as $c = [a, b]$.

$$a = [a_1, a_2, a_3], b = [b_1, b_2\ , b_3]$$

$$c = [a, b] \rightarrow [a_1, a_2, a_3, b_1, b_2, b_3]$$

Another operation is *vector addition.* It is to sum element values of two or more vectors at each index together. This operation operates as:

$$a + b = [a_1 + b_1, \dots, a_n + b_n]$$

The *transpose of a matrix* is to flip a matrix over its main diagonal. The transpose of a matrix $A$ is represented as $A^{\intercal}$. The transpose of the matrix $A$ explicitly illustrates as:

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \end{bmatrix}$$

$$A^{\intercal} = \begin{bmatrix} a_{1,1} & a_{2,1} \\ a_{1,2} & a_{2,2} \\ a_{1,3} & a_{2,3} \end{bmatrix}$$

A vector can be transposed as well. The vector can be represented as a matrix with one column as explained above. The transposition of a vector can be illustrated as:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \to x^{\intercal} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}$$

The next important operation is *matrix-vector product*. This operation is the multiplication between a matrix and a vector. The notation of this matrix-vector product is represented as $Ax$, where $A \in \mathbb{R}^{m \times n}$ is a matrix and $x \in \mathbb{R}^{n}$ is a vector. The result from this operation $Ax$ is $b$, where $b \in \mathbb{R}^{m}$. The matrix-vector product formula is as follows:

$$Ax = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ a_{2,1} & \ldots & a_{2,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{1,1}x_1 & + & \cdots & + & a_{1,n}x_1 \\ a_{2,1}x_2 & + & \ldots & + & a_{2,1}x_2 \\ \vdots & & \ddots & & \vdots \\ a_{m,1}x_n & + & \cdots & + & a_{m,n}x_n \end{bmatrix}$$

## 3.2 Profile-Based Social Media Dataset

Datasets from profile-based social media platforms usually contain a set of users and their profiles. Figure 3-1 depicts a dataset collected from those platforms. Collected users are normally labelled at the *user-level* such as depressed or non-depressed, as can be seen from the smiling and frowning faces. These labels can help to classify between user groups. Each user collected contains a set of posts on their profiles. However, each individual post is not normally labelled and labelling each post is an extensive task. We call a post labelled as *post-level*. A post-level label can clarify what topic and who the post is mentioning to, which can provide more understanding and help to detect depressed users more correctly. With this nature of profile-based social media datasets, we are proposing a predictive model that can classify users with depression and identify posts relevant to mental health topics and self-disclosure of the posters.

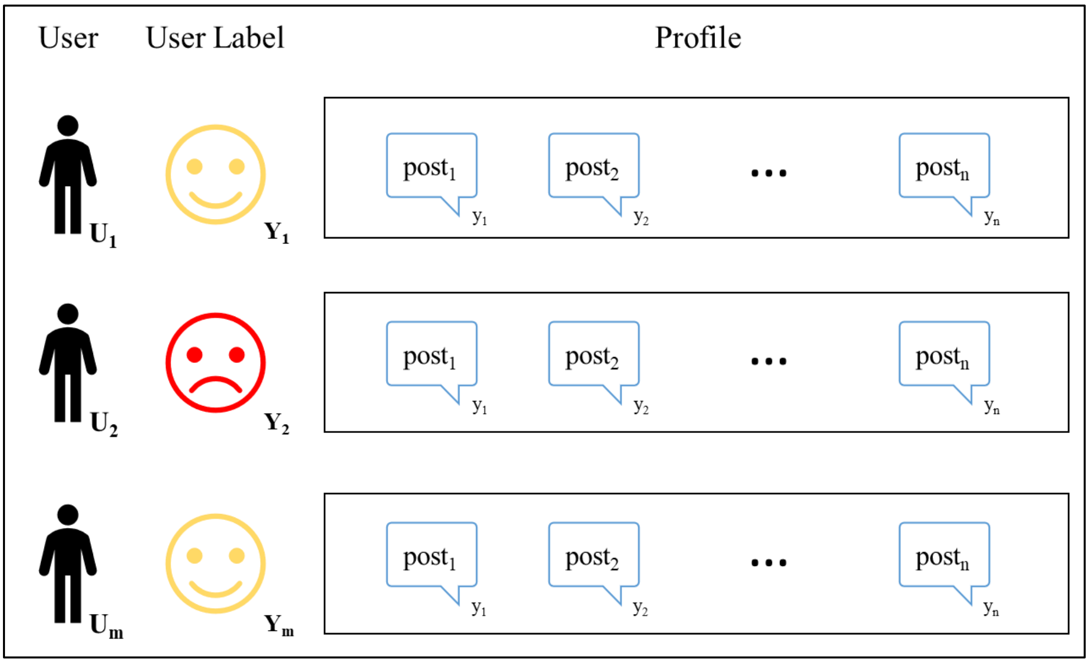


*Figure 3-1 Datasets collected from profile-based social media platforms contain users associated with labels and a set of posts on their profiles.*

To formalise our dataset associated with a MIL problem, a dataset $D$ contains a set of bags (users) $\{U_1, U_2, \dots, U_m\}$, where each user $U_i$ is associated with a label $Y_i$ (depressed/non-depressed) and consists of a set of instances (posts) $\{p_1, p_2, \dots, p_n\}$. $m$ is the number of users in the dataset. $n$ is the number of posts on the user profile $U_i$, which varies from user to user. From these assumptions, it can be said that $U_i = \{p_1, p_2, \dots, p_n\} \in D$ and $p_j \in U_i$. We assume that an individual post $p_j$ should have its own a label $y_j$. To provide this individual label $y_j$, we will develop a predictive model to do this task and attend to self-disclosure of mental symptoms.

## 3.3 Proposed Models

The architecture of our predictive models is based on MIL with neural networks. The basic idea behind MIL is to learn from a set of labelled bags to predict a label of unseen data based on the bags. Instantly, MIL can learn a label of instance transferred from the labelled bags. In this way, the training MIL does not require individual labels for each instance.

The MIL paradigm is suitable for a profile-based social media dataset, since it only had labels for the users but not for their individual posts. Considering the purposes of MIL, it can facilitate the development of a predictive model for detecting social media users with depression and instantly identify posts related to poster’s mental health. Normally, datasets from profile-based social media platforms are labelled at the user-level, but not at the post-level, as interpreted above. This makes it difficult to find which messages posted on social network are related to self-disclosure of mental health disorders. To highlight posts disclosed mental health problems of a user, anaphora resolution with deep learning technique (as explained in appendix A.4.4) will be applied into the MIL.

The proposed architectures of our models are inspired by and follow the hierarchical attention network (HAN) introduced by Yang [121] and the multiple instance learning network (MILNET) proposed by Angelidis [31] and Kotzias [122]. Their models successfully classify web-based review documents and instantly identify the sentiment polarity of each segment of given documents. MILNET comprises segment encoding, segment classification, and document classification via an attention mechanism. Segment encoding transforms sentences in a document into segments via word-embedding matrices and a CNN layer. Each segment representation is classified using a softmax classifier. An attention mechanism based on a bidirectional gated recurrent unit (GRU) is used to weight the important segments to make a final document prediction as the weighted sum of the segment distributions. MILNET performs well in predicting the sentiment of a document and identifying the sentiment of the text segments but is not as successful in identifying the person mentioned in the document.

Given its success, in the present study we adapted and improved the MILNET approach to develop our MIL-SocNet, eliminating the lack of identifying a person mentioned in the document using an anaphora resolution approach. This enables our model to classify users with depression and instantly identify published posts associated with self-disclosure of mental health disorders. Our model also applies an attention layer in both segment classification layer and document classification layer, while the original work used it in the document classification only. These attention layers are designed to help with explainability, which can highlight which words are important in classifying a post related to a health topic and which posts are associated with the sign of depressed users.

This thesis proposes two classifiers: (1) multiple instance learning model without anaphora resolution; and (2) multiple instance learning model with anaphora resolution.

### 3.3.1 Multiple Instance Learning without Anaphora Resolution

Multiple instance learning without anaphora resolution is a predictive model to classify users with depression from their profile-based social media posts without highlighting posts related to self-declaration for mental health issues, so-called *MIL-SocNet*. Our proposed MIL-SocNet architecture consists of post encoder, word attention on a post, post classification, user encoder, post attention, and user classification (see Figure 3-2).

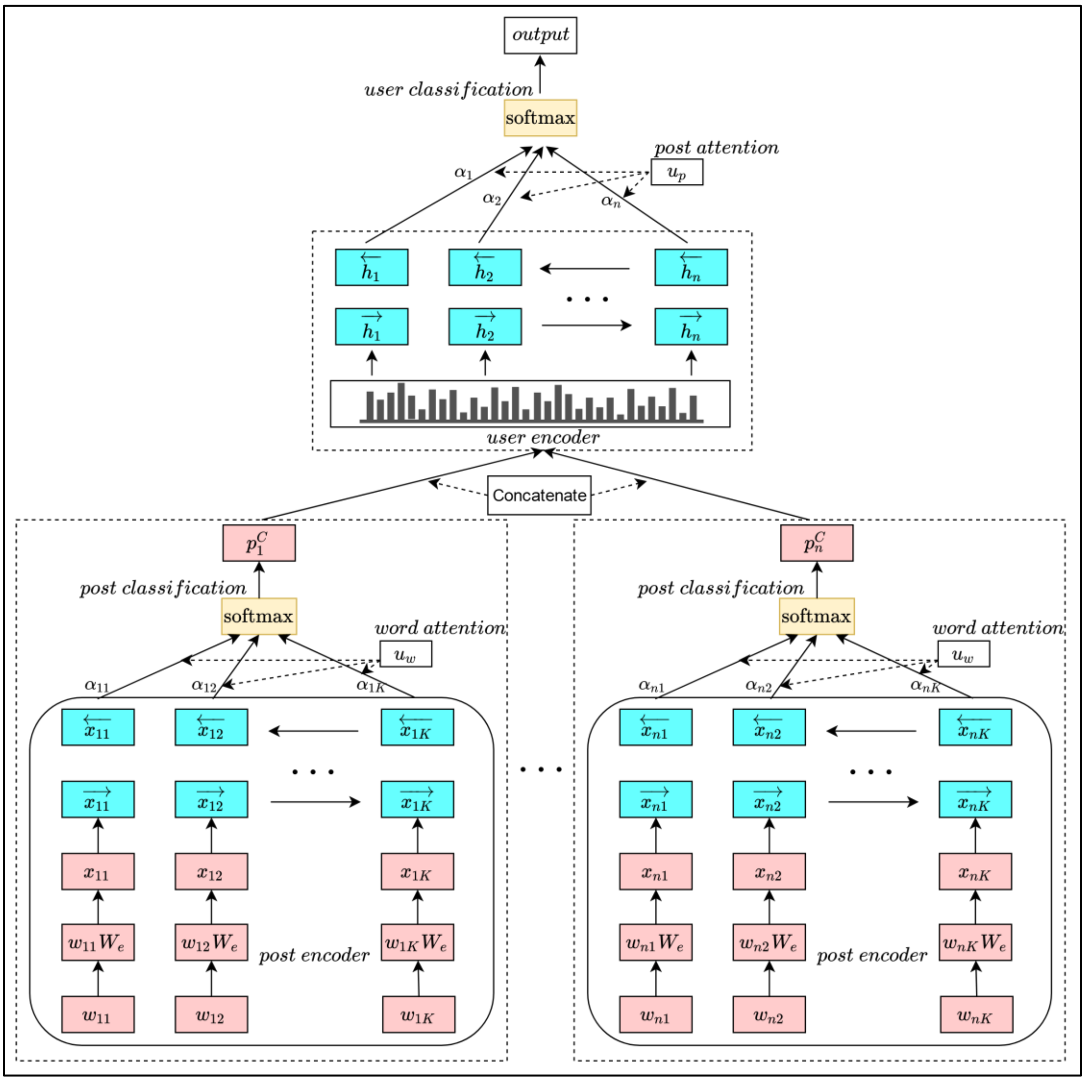


*Figure 3-2. The architecture of our proposed MIL-SocNet*

#### 3.3.1.1 Post Encoder

The first layer of our proposed model transforms each post into machine readable form. First, posts are transformed to word embedding matrices. As mentioned above, an individual user $U_i$ publishes $n$ posts $\{post_1, post_2, \cdots, post_n\}$ and each post contains $K$ words. Note that the size of words $K$ may vary from post to post. Consequently, $w_{ik}$ represents the word $k$ in the $i$-th post $post_i$, where $i \in [1, n]$ and $k \in$

$[1, K]$. $w_{ik}$ is embedded through an embedding matrix $W_e$ to be received a word vector $x_{ik}$. This layer embeds all words $w_{ik}$ of $i$-th post to the word vector:

$$x_{ik} = w_{ik} W_e$$

After embedding all words, a bidirectional LSTM is used to encode the vector:

$$\overrightarrow{h_{\imath k}} = \overrightarrow{LSTM(x_{\imath k})}$$

$$\overleftarrow{h_{\imath k}} = \overleftarrow{LSTM(x_{\imath k})}$$

$$h_{ik} = \left[\overrightarrow{h_{\imath k}}, \overleftarrow{h_{\imath k}}\right], i \in [1, n] \; and \; k \in [1, K]$$

The bidirectional LSTM presents a hidden representation of $h_{ik}$, which is concatenated from the forward hidden state $\overrightarrow{h_{\imath k}}$ and the backward hidden state $\overleftarrow{h_{\imath k}}$ . The word hidden vector is then sent to an attention mechanism to select important words.

### 3.3.1.2 Word Attention on a Post

Not every word is equally representative of a post meaning. An attention mechanism is used to pick the words that best capture the relevance of the post. The attention layer consists of a tanh function to produce an attention vector $u_{ik}$ of the $k$-th word in the $i$-th post, where $W_w \; and \; b_w$ are weights and bias, respectively.

$$u_{ik} = tanh\,(W_w h_{ik} + \; b_w)$$

$$\alpha_{ik} = \frac{\exp\,(u_{ik}^{\intercal} u_w)}{\sum_k \exp\,(u_{ik}^{\intercal} u_w)}$$

The importance of words or attention weights $\alpha_{ik}$ is calculated through the normalised similarity of $u_{ik}$ with the context vector of the word level $u_w$, which is learnt and updated during the training step.

$$p_i = \sum_k \alpha_{ik} \, h_{ik}$$

Finally, the post vector $p_i$ is computed through the weighted sum of word importance $\alpha_{ik}$ with the hidden representation of $h_{ik}$ generated from the bidirectional LSTM.

### 3.3.1.3 Post Classification

To make a prediction about a post related to either a mental health or another topic, each post vector $p_i$ from the attention layer is classified through a softmax function [123].

$$P_i^C = softmax(W_c p_i + \ b_c), C \in [0,1]$$

The function generates the probabilities of post labels $P_i^C = (p_i^1, \cdots, p_i^C)$, where $C \ \in [0, 1]$ with 1 denoting a mental health related post and 0 denoting a non-mental-health related post. Labels used to train this layer are derived and computed from the user level labels only. Parameters $W_c$ and $b_c$ are learnt and updated during the training step. Every predicted post label will be used to teach a predictive model and detect the depressed user $U_i$.

### 3.3.1.4 User Encoder

Detecting users with depression requires a pattern to differentiate between the user groups. To predict those users, this study uses a temporal pattern of posting generated from the post classification layer. This layer concatenates the probabilities of every classified post label into a single list of the label probabilities, so-called "*user representation*" henceforth. The user representations are expected to differ between the two groups. Then, it is passed through a bidirectional LSTM to learn the changing patterns of post categories over observation time. This generates the forward hidden state $\overrightarrow{h_\iota}$ and the backward hidden state $\overleftarrow{h_\iota}$ of the user representation. Finally, they both are concatenated to $h_i$.

$$\overrightarrow{h_\iota} = \overrightarrow{LSTM(P_\iota^C)}$$

$$\overleftarrow{h_\iota} = \overleftarrow{LSTM(P_\iota^C)}$$

$$h_i = [\overrightarrow{h_\iota}, \overleftarrow{h_\iota}], i \in [1, n]$$

#### 3.3.1.5 Post Attention

Not all posts of a user are equally associated with depression. Some posts may contain cues relevant to depression while others may not. For that purpose, an attention mechanism is applied to reward posts that correctly represent the characteristic of the user and are important to correctly detect a user with depression. A one-layer multi-layer perceptron (MLP) produces the attention vector $u_i$ of the $i$-th post. The parameter $W_t$ denotes the weights of the post, and the parameter $b_t$ represents the bias of the post.

$$u_i = \tanh\,(W_t h_i + \ b_t)$$

$$\alpha_i = \frac{\exp\,(u_i^\intercal u_p)}{\sum_i \exp\,(u_i^\intercal u_p)}$$

Attention weights of posts or important posts $\alpha_i$ are computed through the similarity of $u_i$ with the context vector of posts level $u_p$, which is learnt and updated during the training step. The user vector $v$ is achieved through summarising all the information of post label possibilities of the user.

$$v = \sum_{i} \alpha_i \, h_i.$$

#### 3.3.1.6 User Classification

Finally, a predictive model for detecting a user with depression can be achieved through the user vector $v$ derived from encoding the concatenation of the probabilities and the attention weights of the classified post labels from the user. A softmax function is again used to perform the classification.

$$P_U^C = softmax(W_C v + \ b_C)$$

### 3.3.2 Multiple Instance Learning with Anaphora Resolution

We extended the MIL-SocNet model with anaphoric resolution, to create the *MILA-SocNet* model. We present this model to improve performance of predictive power by adding anaphora resolution encoder to ensure the algorithm focuses on posts related to the author or self-disclosure. Our proposed MILA-SocNet architecture consists of post encoder, word attention on a post, post classification, anaphora resolution encoder, user encoder, post attention, and user classification (see Figure 3-3). The details of post encoder, word attention on a post, post classification, user encoder, post attention, and user classification are the same as the MIL-SocNet explained above. This section will describe the details of *anaphora resolution encoder*.

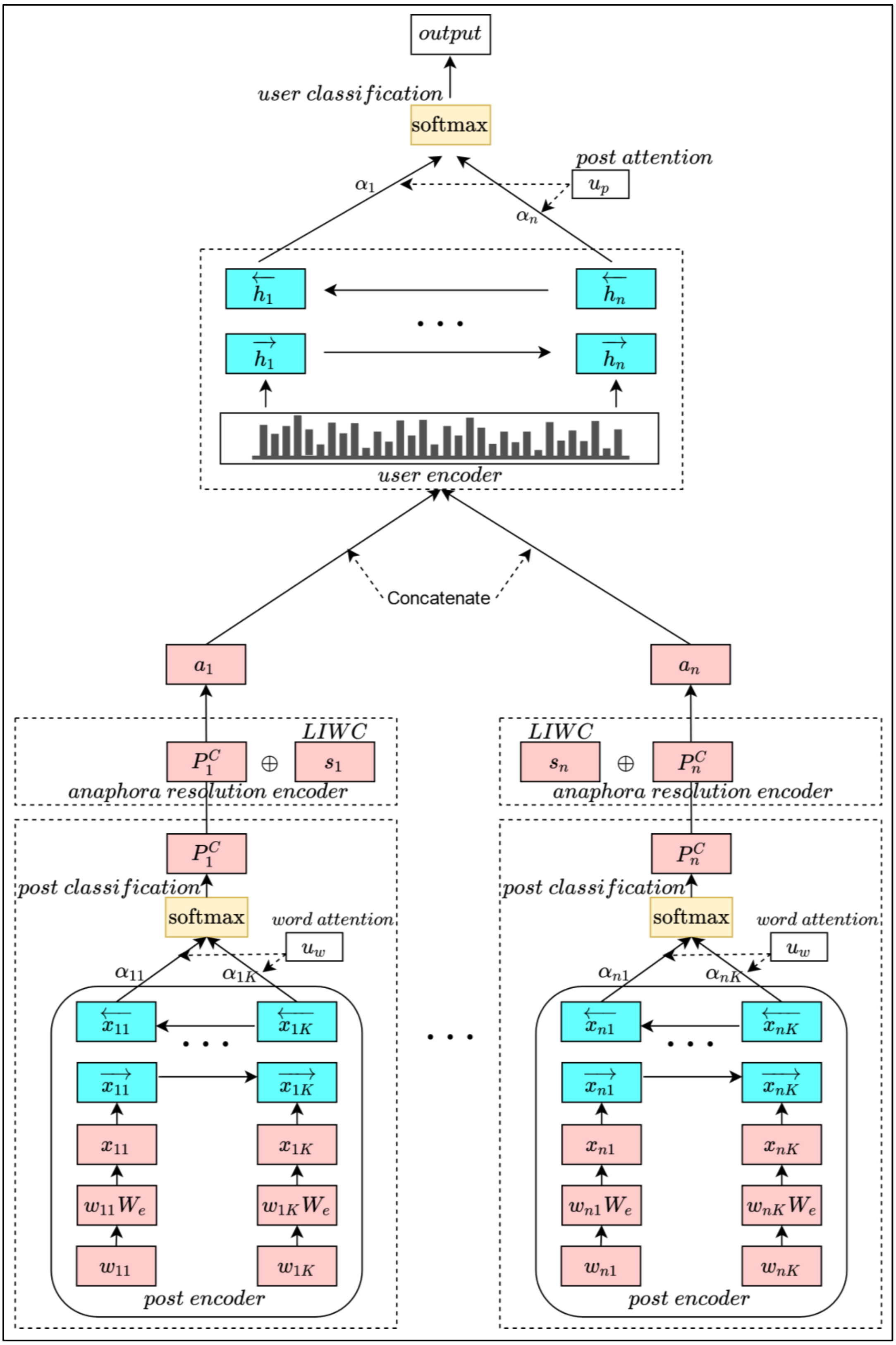


Figure 3-3 The architecture of our proposed MILA-SocNet.

#### 3.3.2.1 Post encoder

According to the above section, the post encoder transforms individual words $w_{ik}$ of the $i$-th post through an embedding matrix $W_e$ to be received a word vector $x_{ik}$. It is then passed through a bidirectional LSTM:

$$x_{ik} = w_{ik}W_e$$

$$\overrightarrow{h_{\imath k}} = \overrightarrow{LSTM(x_{\imath k})}$$

$$\overleftarrow{h_{\imath k}} = \overleftarrow{LSTM(x_{\imath k})}$$

$$h_{ik} = \left[\overrightarrow{h_{\imath k}}, \overleftarrow{h_{\imath k}}\right], i \in [1, n] \; and \; k \in [1, K]$$

This results in the word hidden vector $h_{ik}$.

#### 3.3.2.2 Word Attention on a Post

After receiving the word hidden vector from the post encoder, an attention mechanism is used to pick the most important words in the $i$-th post $p_i$. The mechanism is based on a tanh function:

$$u_{ik} = tanh\,(W_w h_{ik} + b_w)$$

Attention weights $\alpha_{ik}$ are computed as follows:

$$\alpha_{ik} = \frac{\exp\,(u_{ik}^{\intercal} u_w)}{\sum_k \exp\,(u_{ik}^{\intercal} u_w)}$$

Finally, multiplication of the word importance $\alpha_{ik}$ by the hidden representation $h_{ik}$ is summed to be obtained the post vector $p_i$:

$$p_i = \sum_k \alpha_{ik}\, h_{ik}$$

### 3.3.2.3 Post Classification

To receive a predicted label of the $i$-th post $P_i^C$, the post vector $p_i$ is computed through a softmax classifier:

$$P_i^C = softmax(W_c p_i + b_c), C \in [0,1]$$

### 3.3.2.4 Anaphora Resolution Encoder

For the MILA-SocNet model, we added an anaphora resolution encoder to attend to self-interpretation. Pronoun features from LIWC (as explained in *appendix A.4.2*) are used to add informative interpretation to each post being analysed for emotions, thinking styles, social states, parts of speech, and psychological dimensions.

Each post is combined between the extracted pronoun features $s_i$ and the post classified label $P_i^C$ from the post classification layer. This yields an anaphora resolution vector as follows:

$$a_i = [s_i, P_i^C]$$

### 3.3.2.5 User Encoder

Each anaphora resolution vector $a_i$ is concatenated together to create "*user representation*". The concatenated vector is then passed through a bidirectional LSTM to learn the text categories and the anaphoric resolution features. This generates $h_i$ combined from the forward hidden state $\overrightarrow{h_\iota}$ and the backward hidden state $\overleftarrow{h_\iota}$.

$$\overrightarrow{h_\iota} = \overrightarrow{LSTM(a_\iota)}$$

$$\overleftarrow{h_\iota} = \overleftarrow{LSTM(a_\iota)}$$

$$h_i = [\overrightarrow{h_\iota}, \overleftarrow{h_\iota}], i \in [1, n]$$

### 3.3.2.6 Post Attention

Similar to MIL-SocNet, an attention mechanism was used to pay attention to content with self-disclosure of symptoms of mental disorder. Based on MIL-SocNet picking up important posts relevant to mental health issues, MILA-SocNet can pay more attention to posts related to self-disclosure of mental disorder from those additional features in the anaphora resolution encoder. The mechanism consists of a one-layer MLP to produce the attention vector $u_i$:

$$u_i = \tanh\,(W_t h_i + \ b_t)$$

The similarity of $u_i$ with the context vector of posts $u_p$ are computed to be received the importance of the post and get an attention weight $\alpha_i$ through a softmax function. The user vector $v$ is achieved through summarising all the information of the hidden vector of the post $h_i$ with attention weights $\alpha_i$.

$$\alpha_i = \frac{\exp\,(u_i^{\intercal} u_p)}{\sum_i \exp\,(u_i^{\intercal} u_p)}$$

$$v = \sum_i \alpha_i \, h_i.$$

### 3.3.2.7 User Classification

In the end, the user classification is computed through a softmax classifier with the user vector $v$ derived from the post attention layer.

$$P_U^C = softmax(W_C v + \ b_C)$$

## 3.4 Computational Example

After explaining the nature of a profile-based social media dataset and our proposed MIL-SocNet and MILA-SocNet generated models, we will now describe a practical example of how the models work on a dataset.

For example, a user $U_1$ has three posts on his profile. They consist of *$post_1$=I am feeling blue.*; *$post_2$=Today is Friday!!!*; and *$post_3$=I am anxious about my exams*.

First, the post encoder transfers every word of the individual posts $w_{ik}$ into a word vector through a word embedding matrix $W_e$, $x_{ik} = w_{ik}W_e$. To make it easily understandable, we assume that each word in the embedding matrix has a 2-dimensional vector. Then, the embedding matrix is $W_e \in \mathbb{R}^{m \times n}$, where $m$ denotes the number of words in the matrix and $n$ represents the number of dimensional spaces of words.

$$W_e = \begin{bmatrix} i & \rightarrow & 0.2 & 0.3 \\ am & \rightarrow & 0.1 & 0.9 \\ feeling & \rightarrow & 0.3 & 0.5 \\ blue & \rightarrow & 0.8 & 0.7 \\ today & \rightarrow & 0.5 & 0.6 \\ is & \rightarrow & 0.4 & 0.2 \\ & \cdots & & \\ exams & \rightarrow & 0.2 & 0.9 \end{bmatrix}$$

The encoder generates word vectors of $post_1$ through the word embedding matrix as follows:

$$x_{11} = [0.2, 0.3] \rightarrow I$$

$$x_{12} = [0.1, 0.9] \rightarrow am$$

$$x_{13} = [0.3, 0.5] \rightarrow feeling$$

$$x_{14} = [0.8, 0.7] \rightarrow blue$$

Each word vector is passed through a bidirectional LSTM to be received a hidden word vector $h_{ik}$. Assume we compute the word vector $x_{11}$ through the bidirectional LSTM:

$$\overrightarrow{h_{11}} = \overrightarrow{LSTM([0.2, 0.3])} \rightarrow [0.1, 0.3]$$

$$\overleftarrow{h_{11}} = \overleftarrow{LSTM([0.2, 0.3])} \rightarrow [0.8, 0.7]$$

Then, the layer concatenates two hidden word vectors $\overrightarrow{h_{11}}$ and $\overleftarrow{h_{11}}$:

$$h_{11} = \left[\overrightarrow{h_{11}}, \overleftarrow{h_{11}}\right]$$

This results in:

$$h_{11} = [0.1, 0.3, 0.8, 0.7]$$

Each hidden word vector $h_{ik}$ is computed through a word attention layer to produce a post vector $p_i$. First, the layer computes:

$$u_{ik} = tanh(W_w h_{ik} + \ b_w)$$

We will demonstrate the above equation step-by-step to make it easy to understand and follow. This begins with the *matrix-vector product* (as explained above) between a random matrix:

$$W_w = \begin{bmatrix} 0.1 & 0.5 & 0.3 & 0.0 \\ 0.2 & 0.6 & -0.6 & 0.8 \\ 0.3 & -0.7 & -0.9 & 0.6 \\ 0.4 & 0.8 & 0.2 & 0.4 \end{bmatrix}$$

and the vector $h_{11}$. The result from the operation is:

$$W_w h_{11} = \begin{bmatrix} 0.1 & 0.5 & 0.3 & 0.0 \\ 0.2 & 0.6 & -0.6 & 0.8 \\ 0.3 & -0.7 & -0.9 & 0.6 \\ 0.4 & 0.8 & 0.2 & 0.4 \end{bmatrix} \begin{bmatrix} 0.1 \\ 0.3 \\ 0.8 \\ 0.7 \end{bmatrix}$$

$$W_w h_{11} = \begin{bmatrix} 0.1 * 0.1 & + & 0.5 * 0.3 & + & 0.3 * 0.8 & + & 0.0 * 0.7 \\ 0.2 * 0.1 & + & 0.6 * 0.3 & - & 0.6 * 0.8 & + & 0.8 * 0.7 \\ 0.3 * 0.1 & - & 0.7 * 0.3 & - & 0.9 * 0.8 & + & 0.6 * 0.7 \\ 0.4 * 0.1 & + & 0.8 * 0.3 & + & 0.2 * 0.8 & + & 0.4 * 0.7 \end{bmatrix} = \begin{bmatrix} 0.40 \\ 0.28 \\ -0.48 \\ 0.72 \end{bmatrix}$$

The next step is to add the bias vector $b_w = [0.2, 0.1, 0.2, 0.3]$ and the vector $W_w h_{11}$ together:

$$W_w h_{11} + b_w = \begin{bmatrix} 0.40 + 0.2 \\ 0.28 + 0.1 \\ -0.48 + 0.2 \\ 0.72 + 0.3 \end{bmatrix} \rightarrow \begin{bmatrix} 0.60 \\ 0.38 \\ -0.28 \\ 1.02 \end{bmatrix}$$

Then, passing the vector $W_w h_{11} + b_w$ through a tanh function results in:

$$u_{ik} = tanh\left(\begin{bmatrix} 0.60 \\ 0.38 \\ -0.28 \\ 1.02 \end{bmatrix}\right) \rightarrow \begin{bmatrix} 0.54 \\ 0.36 \\ -0.27 \\ 0.77 \end{bmatrix}$$

Next is to compute a softmax function:

$$\alpha_{ik} = \frac{\exp\,(u_{ik}^{\intercal} u_w)}{\sum_k \exp\,(u_{ik}^{\intercal} u_w)}$$

This function performs the transpose of the vector $u_{ik}^{\intercal}$ and then multiplies by the random vector $u_w = [0.1, 0.2, 0.3, 0.4]$. This results in:

$$u_{11}^{\intercal} u_w = [0.54 \quad 0.36 \quad -0.27 \quad 0.77] \begin{bmatrix} 0.1 \\ 0.2 \\ 0.3 \\ 0.4 \end{bmatrix}$$

$$u_{11}^{\intercal} u_w = [0.54 * 0.1 + 0.36 * 0.2 - 0.27 * 0.3 + 0.77 * 0.4]$$

$$u_{11}^{\intercal} u_w = [0.35]$$

We received $u_{11}^{\intercal} u_w = [0.35]$ through the above equation. The next step is to compute a softmax function:

$$\alpha_{11} = \frac{\exp\,(u_{11}^{\intercal} u_w)}{\sum_k \exp\,(u_{1k}^{\intercal} u_w)}, k \in [1{,}4]$$

$$\alpha_{11} = \frac{\exp\,([0.35])}{\sum_k (\exp([0.35]) + \exp([0.91]) + \exp([-0.87]) + \exp\,([0.49]))}, k \in [1{,}4]$$

Let assume we receive the vector $\alpha_{11} = [0.24]$ computed from $u_{11}^{\intercal} u_w = [0.35]$, $u_{12}^{\intercal} u_w = [0.91]$, $u_{13}^{\intercal} u_w = [-0.87]$, and $u_{14}^{\intercal} u_w = [0.49]$. We then compute a weighted sum of $\alpha_{ik}$ and $h_{ik}$:

$$p_i = \sum_k \alpha_{ik}\, h_{ik}, k \in [1, K]$$

$$p_i = \sum_k \left( \left( [0.24] \quad \begin{bmatrix} 0.1 \\ 0.3 \\ 0.1 \\ 0.7 \end{bmatrix} \right) + \left( [0.41] \quad \begin{bmatrix} 0.8 \\ 0.9 \\ 0.4 \\ 0.2 \end{bmatrix} \right) + \left( [0.17] \quad \begin{bmatrix} 0.7 \\ 0.9 \\ 0.6 \\ 0.5 \end{bmatrix} \right) + \left( [0.29] \quad \begin{bmatrix} 0.4 \\ 0.3 \\ 0.8 \\ 0.4 \end{bmatrix} \right) \right)$$

$$p_i = \sum_k \left( \left( \begin{bmatrix} 0.02 \\ 0.07 \\ 0.02 \\ 0.17 \end{bmatrix} \right) + \left( \begin{bmatrix} 0.33 \\ 0.37 \\ 0.16 \\ 0.08 \end{bmatrix} \right) + \left( \begin{bmatrix} 0.12 \\ 0.15 \\ 0.10 \\ 0.09 \end{bmatrix} \right) + \left( \begin{bmatrix} 0.12 \\ 0.09 \\ 0.23 \\ 0.12 \end{bmatrix} \right) \right)$$

$$p_i = \begin{bmatrix} 0.59 \\ 0.68 \\ 0.52 \\ 0.45 \end{bmatrix}$$

Finally, a post classification is computed through a softmax function:

$$P_i^C = softmax(W_c p_i + b_c)$$

$$P_i^C = \frac{\exp\,(W_c p_i + b_c)}{\sum_i \exp\,(W_c p_i + b_c)}$$

We need to compute two classes of mental health-related and another topic. Then, the weight matrix $W_c \in \mathbb{R}^{2\times 4}$ and the bias vector $b_c \in \mathbb{R}^2$, where these two variables are shared across all posts.

$$W_c = \begin{bmatrix} -0.1 & -0.5 & 0.1 & 0.9 \\ 0.2 & 0.6 & -0.3 & 0.5 \end{bmatrix}, b_c = \begin{bmatrix} -0.5 \\ 0.3 \end{bmatrix}$$

Then the result of $W_c p_i + \ b_c$ is:

$$W_c p_1 + \ b_c = \begin{bmatrix} 0.06 \\ 0.60 \end{bmatrix} + \begin{bmatrix} -0.5 \\ 0.3 \end{bmatrix} \rightarrow \begin{bmatrix} -0.44 \\ 0.90 \end{bmatrix}$$

$$P_1^C = \frac{\exp(-0.44)}{\sum_i (\exp(-0.44) + \exp(0.90))}$$

$$P_1^C = [0.21, 0.79]$$

We have explained the most important variables and operations used throughout our proposed models. We omit the rest because the previously mentioned operations repeat.

## 3.5 Model Differences

This section will explain what are the main differences between MIL-SocNet and MILA-SocNet algorithms. As described above, MILA-SocNet includes an additional anaphora resolution encoder to attend to content associated with self-disclosure of mental health issues, but MIL-SocNet does not.

Anaphora resolution can help understand textual content associated with stories of the posters posted on profile-based social media platforms. As we have known so far, there is no study directly applying anaphora resolution to detect users with depression from social media data.

Previous studies on detecting users with mental disorders were not aware of anaphora resolution. These examples illustrate content of the posters focused on other stories following:

- *I can't imagine how people with depression go through. Despite being never diagnosed with depression, I couldn't do anything, when my relative passed away. #BellLetsTaIkday*
- I read an article about one who gave birth and was diagnosed with depression. I believe this is a difficult chapter in her life.

Without applying anaphora resolution, a predictive model would detect those users as depressed labels, because their content seems mentioning about their lives and negative events. When looking into details, the negative events are mentioning one's stories.

We have found some studies on applying anaphora resolution to textual content. Anaphora resolution could significantly improve the performance of opinion mining systems on product/movie reviews [124]. Anaphora resolution has been applied to sentiment analysis, which helps infer references to previous sentences [114]. Recently, anaphora resolution trained on neural networks achieved good performance in coreference resolution (a task of identifying expressions refer to the same entity) [125, 126].

Aktaş et al. [32] studied anaphora resolution from social media data. They manually annotated genres of anaphora resolution for Twitter conversations. Othman et al. [127] also investigated tweet conversations for product feature extraction with considering anaphora resolution importance. Both studies noticed that users used the first (I, we), second (you) and third (he, she, it, they) person/thing pronouns referring to participants or product reviews. Anaphora resolution could help to understand phenomena of content.

Moving to research on detecting depressed users from profile-based social media platforms, studies [22, 24] collected and analysed data from microblogging platforms. They found that users with depression significantly used first, second, and third pronouns. Those users showed high self-attentional focus with frequent first pronoun usage on their social media posts. Similarly, Eichstaedt et al. [28] studied depressed users from a profile-based social network platform and found that the users used first pronouns more often than a normal group.

Consequently, this study focuses on first, second, and third pronoun usage derived from LIWC as an additional set of features to implement our MILA-SocNet. Those additional features are expected to be useful to distinguish between content of self-attentional focus and other’s stories. It is expected that instantly identifying self-attentional content related to mental symptoms can improve performance of the predictive model.

MIL-SocNet and MILA-SocNet algorithms are trained end-to-end on textual content of users with user-level labels. We do not give any additional weights to those features. The models will produce attention weights to important posts and self-disclosure with signs of mental illness by attention mechanism.

## 3.6 Summary

This chapter explained the architectures of our proposed models for detecting depressed users from their profile-based social media data and identifying posts related to self-disclosure of mental health disorders. It also provided the characteristics of a profile-based social media dataset and how our proposed models can be applied to it.

# Chapter 4

# Evaluating MIL-SocNet and MILA-SocNet Algorithms on Profile-Based Social Media Data

This chapter will explain experiments on detecting users with depression from their profile-based social media data using MIL-SocNet and MILA-SocNet algorithms. It begins with data collection and data description from profile-based social media platforms. We will then describe how to train the proposed predictive models with these data. Results from the predictive models will be shown and compared with previously published models.

People use profile-based social media platforms to post content related to themselves, connect friends, and build new relationships. Due to health-related disclosure on the platforms, this kind of content can be useful for researchers to explore health information of the users [128, 129].

This study focuses on two profile-based social media platforms: one for microblogs (*Twitter*) and another for social networks (*Facebook*). These two platforms are picked out as the example platforms because these are the most popular and English-based platforms. The experiments below will be divided into two sections: one for detecting users with depression from microblogs; and one for detecting users with depression from social networks.

## 4.1 Identifying Users with Depression from Microblogging Data

This section explains experiments on a microblogging (Twitter) dataset to test our proposed MILA-SocNet and MIL-SocNet algorithms. It begins with data collection from the microblogging platform and data annotation. Data preprocessing and experiment setup steps will also be explained. Finally, results from the experiments and discussion will be described.

### 4.1.1 Microblog Dataset

The dataset used in this section was gathered from a microblogging platform, especially *Twitter*. Data collection follows the processes elaborately explained in *appendix B.1*. Figure 4-1 illustrates the data collection processes from the microblog. This data collection method includes three steps.

The *first step* is to search for target statuses using a search phrase which contained the statement "*I was diagnosed with depression*" via the search API. We searched the target tweets between January and May 2019. This resulted in 4,892 tweets from 4,545 unique users. The API normally returned tweets that contained all words of the expression. This demanded for a manually screening process to remove non-relevant tweets. After receiving those target tweets, we manually screened to ensure that the tweets do not refer to jokes, quotes, or someone else’s depressive symptoms. User IDs of tweets passing those criteria were kept for downloading all tweets on timelines of the users who created them. After the verification, 2,132 unique users were left in this dataset.

*Another step* is to randomly capture a control group. This group was randomly selected from users who posted tweets between 1st June and 7th June 2019. Users who did not posted in English and from the depressed group were removed from the list of the control group. This resulted in a list of 2,036 users.

Lastly, all tweets on the profiles of the target and control groups were downloaded. This approach does not require human participation, which results in collecting large amounts of data samples. All tweets and users downloaded in this study were set as public and were anonymised.

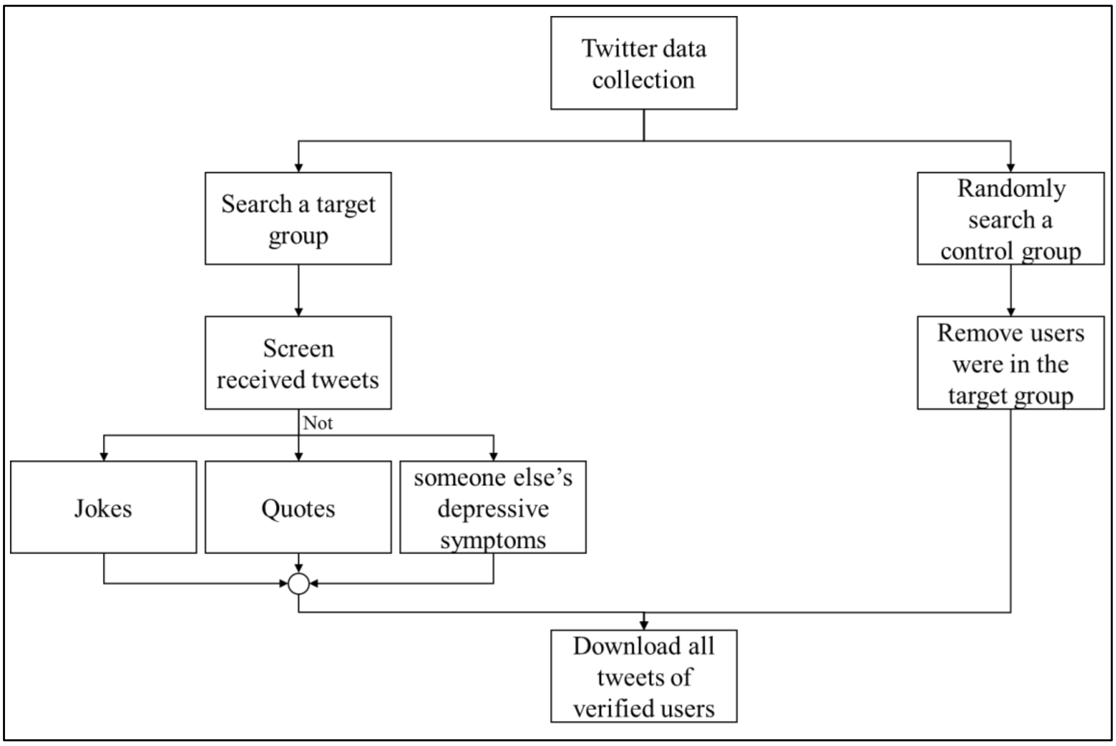

Twitter data collection
Search a target group
Randomly search a control group
Screen received tweets
Remove users were in the target group
Not
Jokes
Quotes
someone else's depressive symptoms
Download all tweets of verified users

*Figure 4-1 The processes of data collection from microblog. First step is search for target tweets and removed non-relevant tweets. Second step is to randomly search control users. Finally, all tweets from target and control groups are downloaded.*

The limits imposed by the Twitter API only allowed us to download the 3,200 most recent Tweets of all verified users from the depressed and control groups. In total, 5m tweets were collected from the 2,132 users with depression and 4.2m tweets from the 2,036 users with no declared depression.

### 4.1.2 Microblog Data Preprocessing

Before developing our MIL model, several transformations were performed. Firstly, the user ID in each message was replaced by a generic “user”. Similarly, any numbers mentioned in messages were replaced by “number” and any specific URLs by “url”. The “#” character in each hashtag was replaced by the string “hashtag” (e.g., “#depression” became “hashtag depression”). Finally, users with fewer than 100 messages and their messages published less than 80% in English (using the pycld2 python library[2]) were removed from the data set, resulting in 3,682 users, 1,983 with declared depression and 1,699 with no declared depression.

All messages in our final dataset were embedded from pre-trained word vectors of GloVe, trained on 2B tweets[3] and supporting 50, 100, and 200 dimensions (as explained in appendix A.4.3). This study used the 100 dimensions variety to transform our messages to word embedding.

[2] https://github.com/aboSamoor/pycld2

[3] http://nlp.stanford.edu/data/glove.twitter.27B.zip

### 4.1.3 Microblog Data Description

The collected dataset was investigated into details to explore differences between the two groups. This reflects the potentials of classifying users with depression from their generated content on the microblog.

Figure 4-2 shows an analysis of data statistics. As can be seen, the proportions of tweets and users between two groups are similar. There were differences between control and depressed users on the number of words per post. Depressed users tended to use more words than the other group. Comparing the ratio of retweets (repost a message published by another user[4]) to tweets, depressed users retweeted more than the control group.

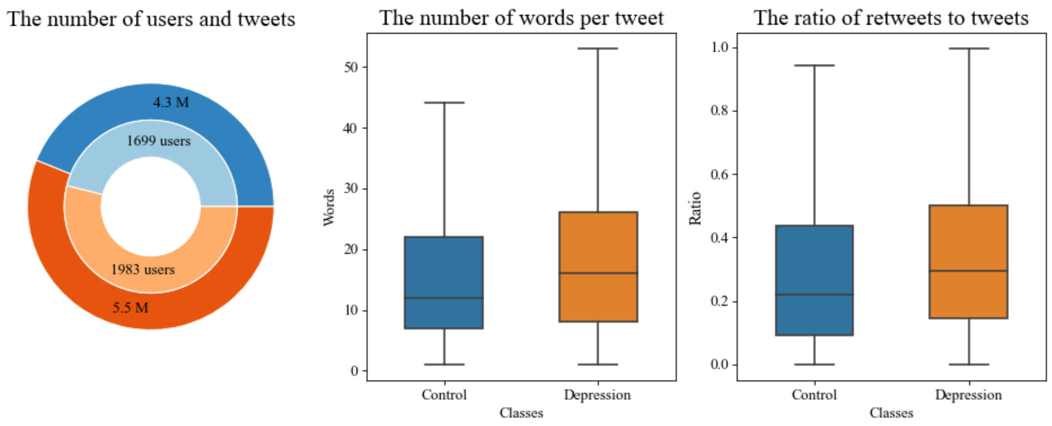


*Figure 4-2 Analysis of data statistics. (Left) shows the percentages of users and tweets between control and depressed users, which the inner circle represents the number of users and the outer circle presents the number of posts. (Middle) demonstrates the number of words per post between two groups. (Right) shows the ratio of retweets to tweets per user between the classes. Blue denotes the control group, and orange is the depressed group.*

[4] https://dictionary.cambridge.org/dictionary/english/retweet

## 4.1.4 Experiment Setup

To train our proposed models, we used the Keras library with TensorFlow backend, a Python library for neural network APIs. An embedding layer of size 100 and an LSTM layer of size 50 were applied in our models. We used an adaptive and momental bound algorithm (AdaMod) [130], and the binary cross-entropy loss function to minimise loss. To prevent overfitting, on each layer a dropout of 0.2 and early stopping were applied to the model during the training step. Batch size of 32 was utilized to train our models in one iteration.

Every message from each user was tokenised, and then the maximum number of tokens was computed. The maximum of 55 tokens were used to preserve all information from the messages, with tweets with fewer than 55 tokens padded with 0 values. The model was trained using 2,000 recent messages from each user, with users with fewer than 2,000 messages having their empty messages padded with 0 values to achieve matching length.

### 4.1.4.1 Baselines and Replicated Models

We also reconstructed a set of published predictive models ranged from classical machine learning to deep learning techniques using user-generated textual content. These models include:

- The LIWC model used LIWC to compute the percentages of words relevant to categories from each post. The extracted percentages were then used to train a predictive model based on a support vector machine with a radial basis function [24]. However, the LIWC model did not have the best performance in the original paper, which used another source to create a corpus and train a model.

We replicated the LIWC model for the convenience of features acquisition from LIWC.

- The Language model was developed from an n-gram, which learns from the sequences of text and computes the probability of unseen text relevant to a category of the train model. This model scored the probabilities of depressed users based on a higher probability of the positive class language model trained from the posts of depressed users or the negative class language model developed from the posts of control users [22].
- The Topic model used LDA to extract topics from text. All posts from each user were used to compute 200 topics, which were then used to develop a logistic regression model for classifying the users with depression [28]. The difference between our experiment and the original work is the hyperparameter execution. We perform a hyperparameter.
- The Usr2Vec model transformed text into an embedding matrix, where words commonly used together were represented in closely dimension spaces, to classify users. The embeddings were learned from posts of users and then summarised as a user representation. The embedding matrices were used to train a predictive model with a classical machine learning technique [61].
- The Deep learning model used word embeddings to represent sequential words of posts of users. A predictive model was trained with a one-dimensional CNN and a global maxpooling layer [62].

### 4.1.5 Results

The performance of every proposed and replicated model for predicting whether each microblog user was likely to be depressed was assessed with cross validation. We split depressive users into 4 chunks and trained the models against all control users. So, each round used 496 depressive users (22.60%) and 1,699 control users (77.40%), which mirrors the real-world incidence of depression. From the total users used in each round, 20% were used as test sets to evaluate the performance of the models. To reserve the same proportions of classes between training and test tests, stratified cross validation was used. Figure 4-3 shows the cross-validation processes.

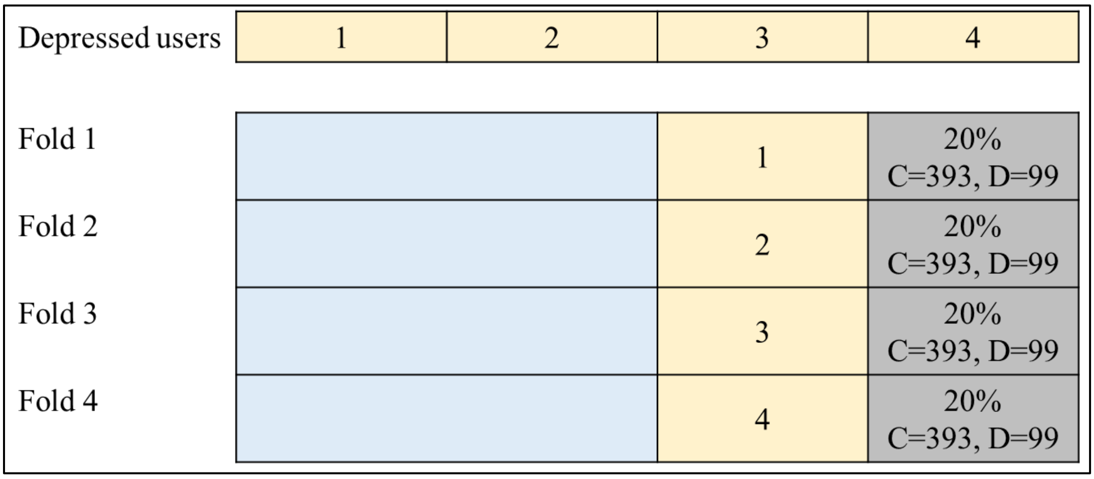


*Figure 4-3 Cross validation on Twitter experiment. C denotes control users and D represents depressed users. Blue, yellow and grey are control data, chunks of depressed users and test sets, respectively.*

Accuracy, precision, recall, and F1-scores were averaged across testing sets. Each of our models and replicated models was trained and tested with the same samples in each fold. To evaluate the performance of our proposed MIL model, we applied several published models (as explained in *section 4.1.4*) to the same data set.

Table 4-1 shows the performance of MILA-SocNet and MIL-SocNet against the alternative models. As can be seen, MILA-SocNet achieves the maximum accuracy (92%), precision (92%), recall (92%), and F1-score (92%) across others. MIL-SocNet yields accuracy, precision, recall, and F1-score of 90%, 91%, 90%, and 90%, respectively. Each model was evaluated with the AUC of a ROC curve. As can be seen in Figure 4-4, MILA-SocNet and MIL-SocNet achieve the highest AUC at 93% and 93%, respectively. Of note, in the present data set, a theoretical model that always predicted the majority class would achieve 77% accuracy. As can be seen, all the models included in Table 6, perform well above this baseline.

*Table 4-1 Performance of our proposed MILA-SocNet and MIL-SocNet models and all replicated models from microblog data. The bold text represents the best result in each metric.*

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
|---|---|---|---|---|
| MILA-SocNet | 92.14% | 0.92 | 0.92 | 0.92 |
| MIL-SocNet | 90.49% | 0.91 | 0.90 | 0.90 |
| Deep learning | 89.07% | 0.89 | 0.89 | 0.89 |
| Usr2Vec | 84.38% | 0.84 | 0.84 | 0.83 |
| LIWC | 83.31% | 0.83 | 0.83 | 0.81 |
| Language | 81.61% | 0.80 | 0.82 | 0.79 |
| Topic | 80.13% | 0.78 | 0.80 | 0.78 |

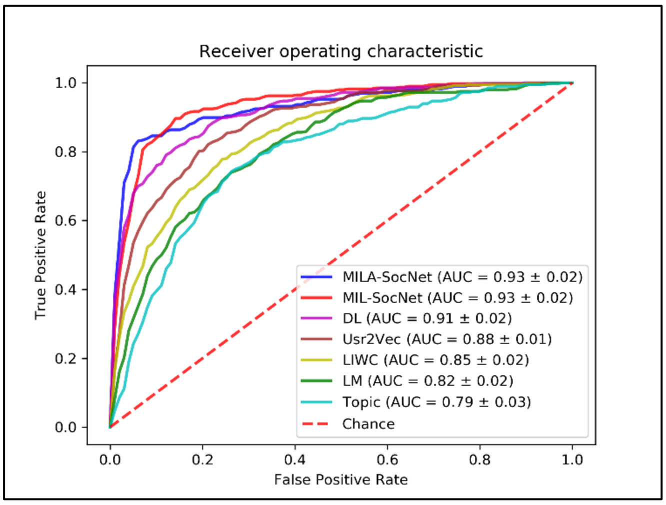


*Figure 4-4. ROC curves of each model. AUCs with SDs of each model are denoted by different colours. The x-axis shows false positive rate and the y-axis presents true positive rate. The dashed line indicates random guess.*

To further compare the performance of MILA-SocNet and MIL-SocNet with other models, AIC (explained in *appendix A.7.5*) was applied across all models. Table 4-2 shows the AICs for each model. Likelihood was computed from the model-based probabilities of observed labels. The number of parameters of the MILA-SocNet, MIL-SocNet and the deep learning models were recovered from the number of trainable parameters reported by the Keras library. The number of parameters of the language model were taken from the number of vocabularies in the positive and the negative language models. The number of parameters of LIWC, Usr2Vec, and topic models were features in the models. The likelihood and AIC were averaged from cross validation as explained above. As can be seen, our model achieves the lowest AIC, which reflects the best performance.

*Table 4-2 The AIC results against all models from a microblog dataset. Each row is reported with the number of parameters (K), likelihood, and AIC. Lower AIC indicates better performance.*

| Model | K | Likelihood | AIC |
|---|---|---|---|
| MILA-SocNet | 59,668 | -143.72 | -597.05 |
| MIL-SocNet | 56,296 | -210.22 | -464.45 |
| Deep learning | 138,502 | -309.97 | -260.84 |
| Language | 16695.5 | -420.31 | -61.03 |
| LIWC | 93 | -169.62 | 575.92 |
| Usr2Vec | 100 | -190.28 | 640.32 |
| Topic | 200 | -276.42 | 1,290.66 |

#### 4.1.5.1 Model Trained with Different Parameters

To further explore our proposed models and compare their performance, a set of different parameters were used to train the models. In the following analyses, the different numbers of posts of each user used to train a model ranged from 500 to 3,200 posts. The numbers of embedded dimensions are 50 and 100. The different lengths of word tokens are the maximum tokens (55) and the average tokens (18). Due to computing limitation, the experiment with post length of 3200 with 100 embedded dimensions and 55 token lengths was not performed. Table 4-3 and Figure 4-5 show the predictive results of MILA-SocNet and MIL-SocNet with different parameters. As can be seen, longer post length and longer word token provides better results, which is to be expected as these provide more textual content. Furthermore, models with fewer embedded dimensions perform worse than those with more dimensions.

*Table 4-3 Performance of our proposed MILA-SocNet and MIL-SocNet models with different parameters. The first number in the model name (first column) represents the number of posts, the second is the number of embedded dimensions, and the last is the number of word token.*

| **Model** | **MILA-SocNet models** | | | | **MIL-SocNet models** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Accuracy** | **Precision** | **Recall** | **F1-score** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| 2000-100-55 | 92.14% | 0.92 | 0.92 | 0.92 | 90.49% | 0.91 | 0.90 | 0.90 |
| 500-100-55 | 85.88% | 0.86 | 0.86 | 0.84 | 84.05% | 0.83 | 0.84 | 0.83 |
| 3200-100-18 | 87.81% | 0.87 | 0.88 | 0.88 | 86.10% | 0.85 | 0.86 | 0.86 |
| 2000-100-18 | 86.90% | 0.86 | 0.87 | 0.86 | 85.65% | 0.85 | 0.86 | 0.85 |
| 500-100-18 | 83.20% | 0.82 | 0.83 | 0.82 | 83.31% | 0.83 | 0.83 | 0.81 |
| 2000-50-18 | 86.62% | 0.86 | 0.87 | 0.86 | 85.42% | 0.85 | 0.85 | 0.85 |
| 500-50-18 | 83.88% | 0.83 | 0.84 | 0.83 | 83.26% | 0.83 | 0.83 | 0.82 |

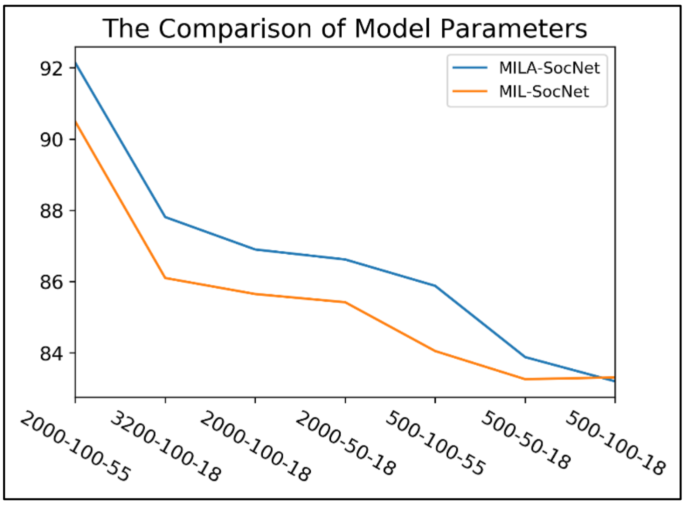


*Figure 4-5 Results from different model parameters. Y axis is the accuracy of the models. X axis represents the number of posts, embedded dimensions, and post tokens in each model, respectively.*

### 4.1.5.2 Explainability and Interpretability of the Models

After the training step, we looked into the explainability and interpretability of MILA-SocNet by observing the attention weights to find out which messages the model paid most attention to.

Two users from each group were randomly selected from those correctly labelled by MILA-SocNet, and attention weights were extracted from the post attention layer. Table 4-4 highlights the messages that achieved the highest and lowest weights for these four users, offering some insight into the model decision-making. It can be seen that our predictive model with anaphoric resolution can identify messages that relate to the posters’ own experiences.

*Table 4-4 Textual content related to self-disclosure of mental illness and attention weights of microblog posts from MILA-SocNet. The text was paraphrased to anonymise users' identities.*

| **Depressed users** | |
|---|---|
| **User 1**<br>***Highest weight***: I was also dealing with depression and anxiety badly. School was hell.<br>***Lowest weight***: Retweet users: Exam without someone's supervision is bad. | **User 2**<br>***Highest weight***: I get some rest, take medication, and engage with what I like. These help me and I do not force myself to do things.<br>***Lowest weight***: Talk about offensive things to physical harm: url. |
| **Non-depressed users** | |
| **User 1**<br>***Highest weight***: The lady Christmas jumper: url.<br>***Lowest weight***: All the best for your match and hope to see you play. | **User 2**<br>***Highest weight***: He reminds me someone in a football team. He can play many positions and he is our best player.<br>***Lowest weight***: People believe you when you have evidence. |

## 4.2 Detecting Users with Depression from Social Network Data

This section will report similar experiments on data from a social network platform. The previous section carried out experiments on a microblog platform and showed that our proposed models could perform well on the dataset. This is experimenting on a different dataset from a different type of social media platform to prove that our proposed models can generalise to profile-based social media datasets. It begins presenting the dataset used in this study. Experiment setup and used parameters will be provided. Finally, it will show and discuss results from the experiments.

### 4.2.1 Social Network Dataset

An example of profile-based social network platforms used in this study is *Facebook* due to the most popular profile- and English-based social network platform. This study used the dataset taken from *myPersonality* [131]. Appendix B.2 explains in detail our data collection tool developed after myPersonality. It provided different psychological results from participants and some of their anonymized Facebook profiles [107]. Detailed information of how to collect data and how to screen participants associated with depression will be explained below.

Participants were recruited online and requested to submit a questionnaire to screen their depressive symptoms. Some participants also provided consent to access their social network profiles within an application designed from the research study.

Figure 4-6 depicts the whole data collection processes from Facebook. First, an app was designed to comply with Facebook regulations. The app must be submitted for the app review and business/individual verification before release the approved app for public. After receiving the app approval, participants are recruited to take part in a research study. The app asks the participants to provide consent to enable profile access and submit a designed questionnaire to screen the participants.

After the app users have provided the permission, a developer can access user's profiles and download their messages on the accessible profiles. The downloaded messages will be kept for research purposes.

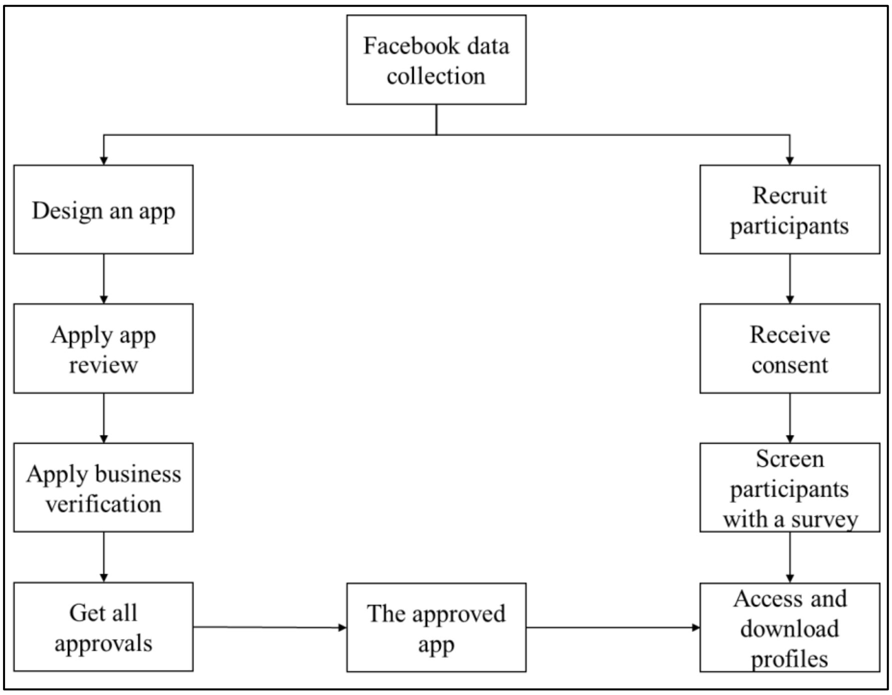


*Figure 4-6 The processes of data collection on Facebook. First step designs and develops an app for review processes. The approved app can be released for Facebook users to use. Last step is to recruit participants and download their Facebook profiles.*

### 4.2.2 Social Network Data Description

The study [131] recruited over 6 million participants through the mypersonality web application. They were asked to take a set of psychometric questionnaires and receive results based on their scores as rewards. The unique participants comprised diverse age groups, backgrounds, and cultures. In our study, we used the *myPersonality* dataset and selected only 5,947 unique participants taking the CES-D questionnaire more than 6,500 times. Participants giving a missing answer in their responses were excluded from our statistical analysis. This resulted in a final sample of 5,291 participants. Table 4-5 shows statistics of participants. 931 of them agreed to share their Facebook profiles. Some of them also provided their gender and age. As shown in the table, both groups had similar statistics, but they differed in gender.

*Table 4-5 Social network data description of participants taking CES-D from myPersonality*

| | **All unique respondents** | **Unique respondents shared profile** |
|---|---|---|
| Number of participants | 5,291 | 931 |
| Female/Male | 3,168/1,825<br>from 4,993 | 551/378<br>from 929 |
| Median age | 22 | 22 |
| Average age | 25.46 (SD= 10.34)<br>from 2,728 | 25.28 (SD= 10.93)<br>from 903 |

### 4.2.3 Social Network Data Preprocessing

Text preprocessing steps were executed on the dataset before training the models. Whitespaces were removed from every post. Any numbers mentioned in posts were replaced by “number”, and any specific URLs by “url”. Posts published less than 80% in English were removed from our dataset. After the processing steps, this resulted in 861 users left.

Every user was labelled as depressed or non-depressed. This study used CES-D (explained in *appendix A.2.3.1*) to classify users. 861 users left from the text preprocessing steps were computed their CES-D scores from responded answers using a cut-off score of 22 to classify users as depressed. This resulted in 294 users with non-depression and 567 users with depression.

Finally, users with fewer than 100 posts were removed from our dataset, resulting in 485 users, 325 with self-reported depression and 160 without depression. A threshold of 100 was used to include users in the training and test sets, because it is the mean number of posts per user, as shown in Figure 4-7 on the right-hand side. There are 45,000 posts from control users and 107,000 posts from depressed users. The average number of words per posts is approximately 21, as shown in Figure 4-7 in the middle.

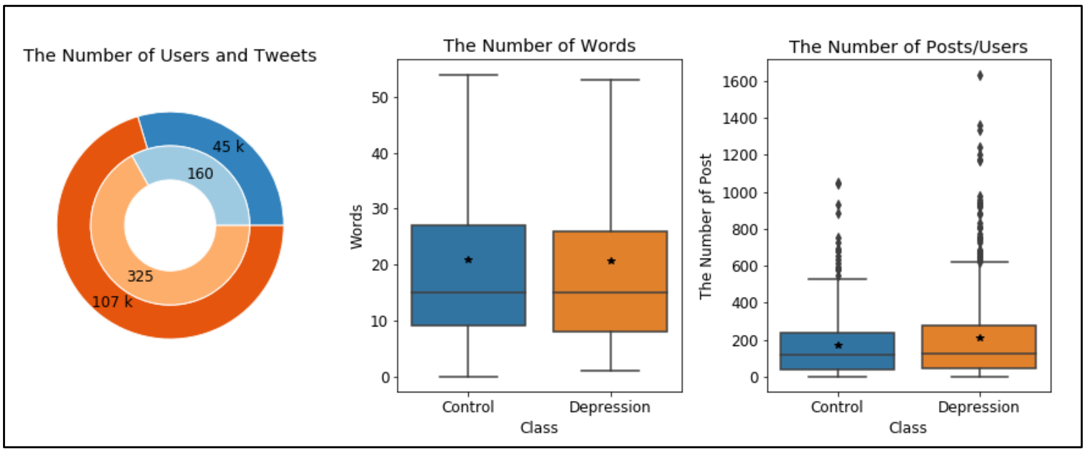


*Figure 4-7 Analysis of data statistics. (Left) shows the percentages of users and posts between control and depressed users, which the inner circle represents the number of users and the outer circle presents the number of posts. (Middle) demonstrates the number of words per post between two groups. (Right) shows the number of posts between the classes. Blue denotes the control group, and orange is the depressed group. Asterisk represents the mean of data.*

### 4.2.4 Experiment Setup

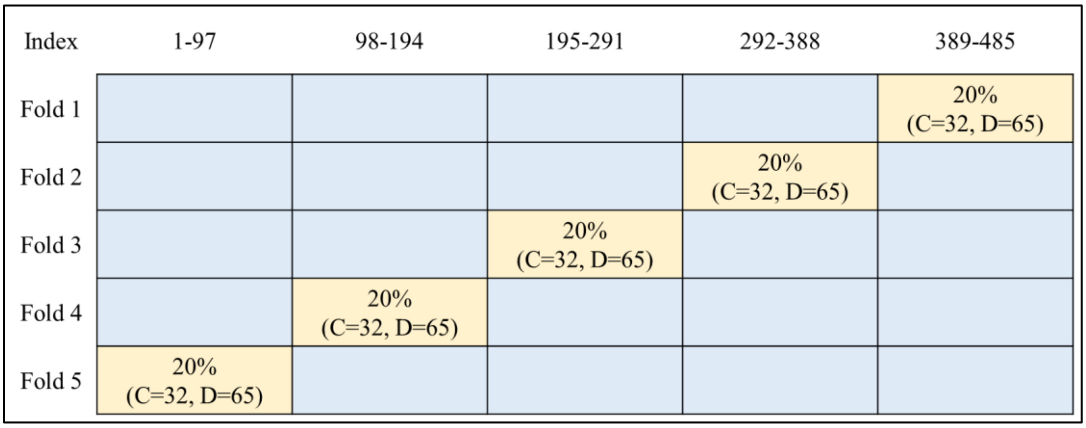
The Keras and Tensorflow libraries were again used to execute our proposed models. Our models used an LSTM layer of size 50. AdaMod and the binary cross-entropy loss function were used to minimise loss and maximise accuracy during training. A dropout of 0.2 and early stopping learning were used to avoid overfitting. Batch size of 32 was utilized to train our models in one iteration.

Every post from each user was tokenised, and it was limited to 44 tokens. 316 recent posts from each user were used to train our proposed models. 0 padding was applied, in the case of tokens and posts less than our above threshold. All tokens were embedded with 100 dimensions from pre-trained word vectors of GloVe, trained on 2B tweets.

### 4.2.5 Results

This section reports the performance of MILA-SocNet and MIL-SocNet on a social network dataset. The models were compared against a set of highly cited models ranged from regression to CNNs by using user-generated text, as explained in *section 4.1.4.1*.

To train and test our proposed and other replicated models, 5-fold cross validation was used. This experiment used a slightly different method for splitting data from the last section. This is because the previous section had a relatively equal number of samples between classes, but this one had a large difference. The previous experiment used chunks of depressed users to test against all control users. This used different chunks of samples to test the models. 20% of the dataset in each fold was reserved as test sets. As explained above, there were 485 users in our final dataset, 325 with self-reported depression (32.99%) and 160 without depression (67.01%). Stratified cross validation was used to reserve the same proportions of classes while splitting data. Figure 4-8 illustrates the processes of the cross validation.



| Index | 1-97 | 98-194 | 195-291 | 292-388 | 389-485 |
|---|---|---|---|---|---|
| Fold 1 | | | | | 20%<br>(C=32, D=65) |
| Fold 2 | | | | 20%<br>(C=32, D=65) | |
| Fold 3 | | | 20%<br>(C=32, D=65) | | |
| Fold 4 | | 20%<br>(C=32, D=65) | | | |
| Fold 5 | 20%<br>(C=32, D=65) | | | | |

*Figure 4-8 Cross validation on Facebook experiment. C denotes control users and D represents depressed users. Blue is training data and yellow is test data.*

Accuracy, precision, recall, and F1-scores were computed to compare performance of our proposed MILA-SocNet and MIL-SocNet models with previously published models. Table 4-6 reports the average results across testing sets from each model. MILA-SocNet achieves the highest accuracy up to 73%, precision 74%, recall 73%, and F1-score 68%. As can be seen, MILA-SocNet achieves the best performance in all dimensions, except precision. Nevertheless, MIL-SocNet still receives the highest precision up to 75% with 72% accuracy, 72% recall, and 67% F1-score. Noted that the majority class of our dataset is 67%. This highlights that our models and all other replicated models achieve results higher than a guessing point line. To ensure that models achieving higher random guessing, ROC curves were plotted on Figure 4-9. As can be seen, MILA-SocNet and MIL-SocNet yield AUCs at 62% and 59%, respectively.

*Table 4-6 Performance of our proposed MILA-SocNet and MIL-SocNet models and all replicated models from social network data. The bold text represents the best result in each metric.*

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| MILA-SocNet | **72.99%** | 0.74 | **0.73** | **0.68** |
| MIL-SocNet | 72.37% | **0.75** | 0.72 | 0.67 |
| Deep learning | 69.48% | 0.69 | 0.69 | 0.64 |
| Language | 70.10% | 0.68 | 0.70 | 0.65 |
| Usr2Vec | 69.28% | 0.67 | 0.69 | 0.64 |
| LIWC | 69.07% | 0.69 | 0.69 | 0.63 |
| Topic | 68.45% | 0.62 | 0.68 | 0.62 |

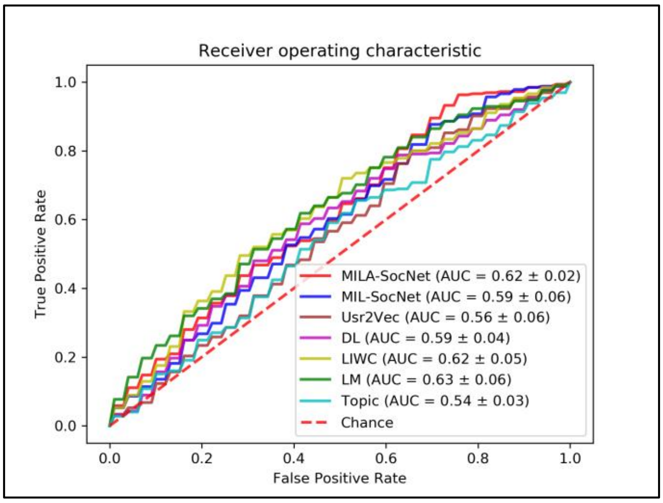


*Figure 4-9 ROC curves of each model. AUCs with SDs of each model are denoted by different colours. The x-axis shows false positive rate, and the y-axis presents true positive rate. The dashed line indicates random guess.*

After reporting the performance of our models and the other replicated models, AICs were computed to compare and select the best model. As mentioned in the previous section, Likelihood was computed from the differences between model-based probabilities and observed labels. The number of parameters of our models and the deep learning models were received from the number of trainable parameters. The number of vocabularies in the language model was used as the number of parameters of the model. Features of LIWC, Usr2Vec, and topic models were used as the number of their parameters. Table 4-7 reports AIC results with across all models. As can be seen, MILA-SocNet achieves the best AIC.

*Table 4-7 The AIC results against all models from social network data. Each row is reported with the number of parameters (K), likelihood, and AIC. Lower AIC indicates better performance.*

| **Model** | **K** | **Likelihood** | **AIC** |
|---|---|---|---|
| MILA-SocNet | 57,204 | -96.12 | -4,872.37 |
| MIL-SocNet | 56,296 | -147.75 | -4,772.72 |
| Deep learning | 138,502 | -64.72 | -4,806.96 |
| Language | 16695.5 | -73.12 | 559.20 |
| LIWC | 93 | -59.98 | 313.46 |
| Usr2Vec | 200 | -65.47 | 339.64 |
| Topic | 200 | -184.25 | 804.66 |

#### 4.2.5.1 Models Trained with Different Parameters

After training all models, we tested our MILA-SocNet and MIL-SocNet with different parameters to find out which set of parameters can achieve the best result. The models were trained under three sets of parameters. Firstly, the number of recent posts was set to 316 (the mean of posts per user) and 647 (90% quantile of the number). Secondly, the number of embedded dimensions were set from 50 to 100. Lastly, the number of word tokens per post is covered 21 (the mean of word tokens from all posts) and 44 (90% quantile of the number). Table 4-8 shows that the best model was trained by 316 recent posts, 100 embedded dimensions, and 44 word tokens. We investigated in more details. Figure 4-10 shows that MILA-SocNet tends to achieve better results than MIL-SocNet. Training models with higher embedded dimensions can yield higher accuracy, as can be seen from Figure 4-11. Figure 4-12 highlights that using the higher number of word tokens can improve accuracy of models.

*Table 4-8 Performance of our proposed MILA-SocNet and MIL-SocNet models with different parameters. The P in the model name (first column) represents the number of posts, the second is the number of embedded dimensions, and the last is the number of word tokens.*

| **Model** | **MILA-SocNet models** | | | | **MIL-SocNet models** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Accuracy** | **Precision** | **Recall** | **F1-score** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| P316-D100-T44 | 72.99% | 0.74 | 0.73 | 0.68 | 72.37% | 0.75 | 0.72 | 0.67 |
| P316-D100-T21 | 72.99% | 0.73 | 0.73 | 0.70 | 70.93% | 0.69 | 0.71 | 0.68 |
| P316-D50-T44 | 71.96% | 0.73 | 0.72 | 0.67 | 71.55% | 0.73 | 0.72 | 0.65 |
| P316-D50-T21 | 71.75% | 0.73 | 0.72 | 0.67 | 71.96% | 0.73 | 0.72 | 0.69 |
| P647-D100-T44 | 70.72% | 0.70 | 0.71 | 0.67 | 70.10% | 0.64 | 0.70 | 0.64 |
| P647-D100-T21 | 72.16% | 0.75 | 0.72 | 0.66 | 69.90% | 0.69 | 0.70 | 0.66 |
| P647-D50-T44 | 71.75% | 0.72 | 0.72 | 0.66 | 70.31% | 0.72 | 0.70 | 0.65 |
| P647-D50-T21 | 70.10% | 0.68 | 0.70 | 0.65 | 71.13% | 0.72 | 0.71 | 0.66 |

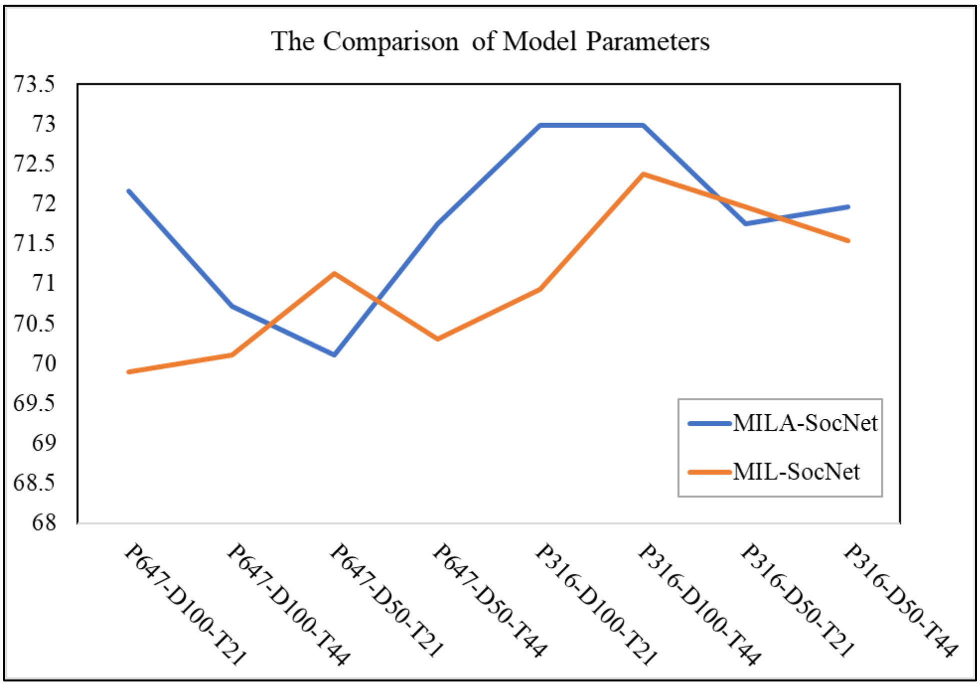


*Figure 4-10 Results from different model parameters. Y axis is the accuracy of the models. X axis represents the number of posts (P), embedded dimensions (D), and post tokens (T).*

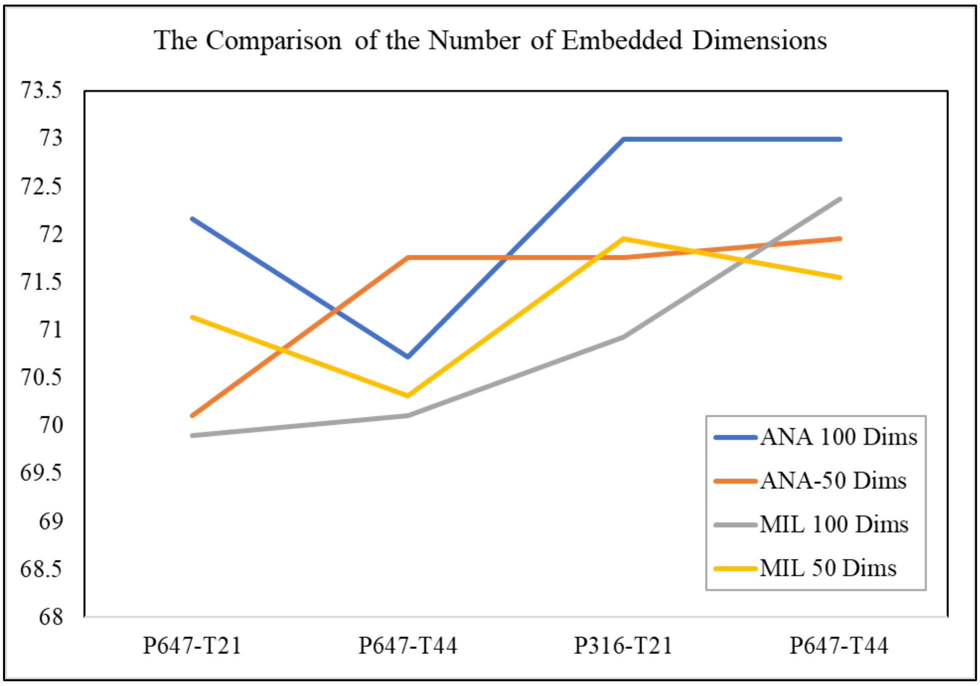


*Figure 4-11 The comparison results from different embedded dimensions. Y axis is the accuracy of the models. X axis represents the number of posts (P) and post tokens (T).*

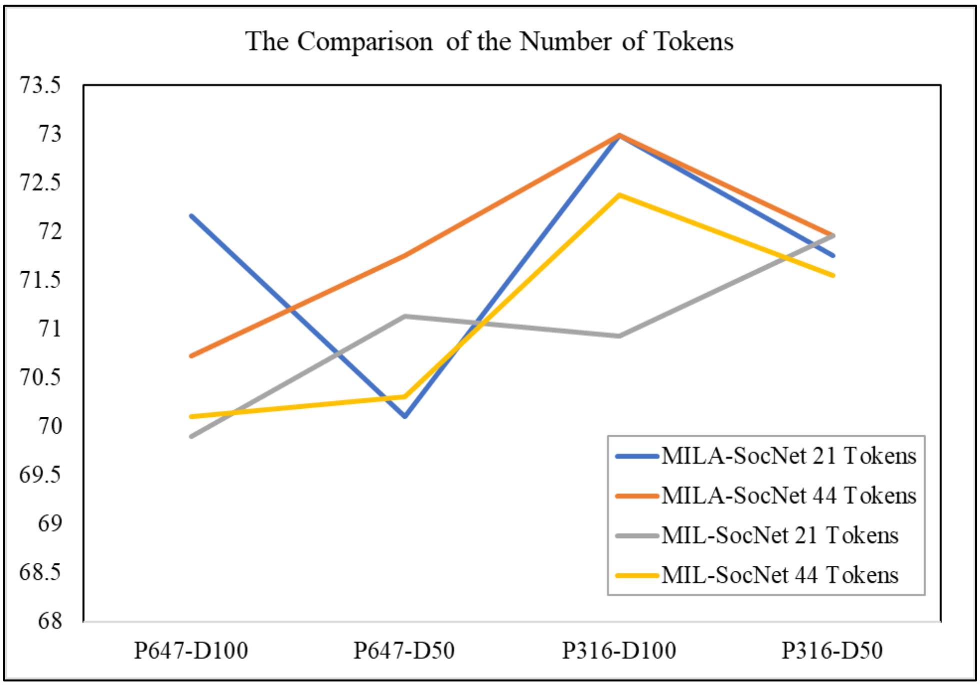


*Figure 4-12 The comparison result from the different number of word tokens. Y axis is the accuracy of the models. X axis represents the number of posts (P) and embedded dimensions (D).*

The results from Table 4-8 show a surprising pattern. We noticed that using a higher number of recent posts does not improve the performance of models. A further investigation was carried out to ascertain what caused this issue. As shown in Figure 4-13, we found that most of the users had a number of posts around the mean (316). Around 170 users posted over 316 status updates. This reflects less than a half of all users had over 316 posts. This could be the reason that the models trained with the higher number of posts did not have enough data to learn. The majority users had fewer than 316 posts.

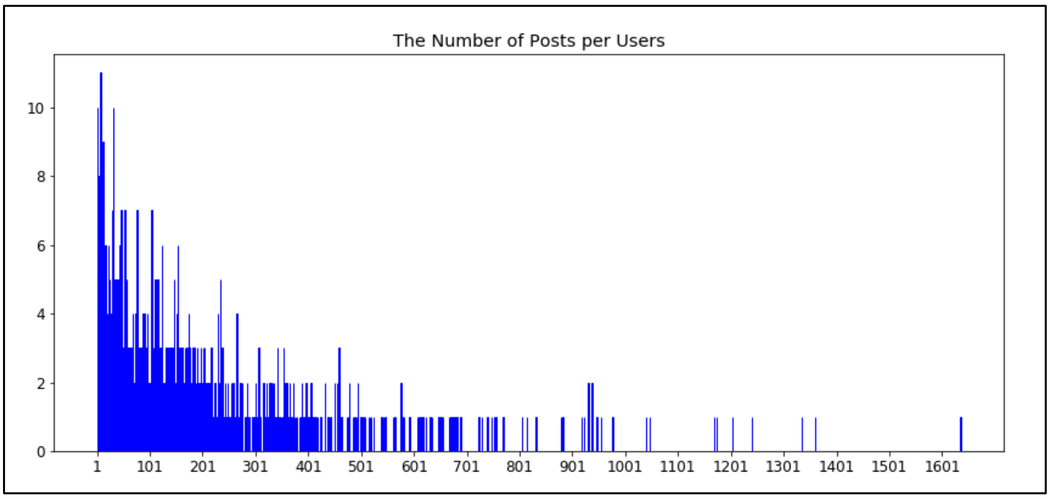


*Figure 4-13 The number of posts per users. X-axis presents the number of posts and y-axis denotes the number of users.*

### 4.2.5.2 Explainability and Interpretability of the Models

To test the explainability and interpretability of MILA-SocNet from the social network dataset, we extracted attention weights from post attention layer (defined in *section 3.3.1.5*). The layer can provide information about posts relevant to mental health problems and posters' own experiences.

We looked for users correctly labelled by the model and extracted information of those posts. Table 4-9 shows the highest and lowest attention weights from 2 accurately labelled users with depression and 2 without depression.

As can be seen from the depressed users, our model could provide posts related to symptoms of depression such as *broken heart* (the depressed user 1) and *some physical impairment* (the user 2). The model paid less attention to posts other than mental health-related and other's own experiences. MILA-SocNet did not specify any post associated with symptoms of mental illness from the control users. This highlights that the post attention layer can detect posts relevant to mental health problems and users own experiences.

*Table 4-9 Textual content related to self-disclosure of mental illness and attention weights of social network posts from MILA-SocNet. The text was paraphrased to anonymise users' identities.*

| **Depressed users** | |
|---|---|
| **<u>User 1</u>**<br>***Highest weight*: I miss you very much. To think of it, it hurts my heart.<br>***Lowest weight*: Finished with season one of an animated series! Keep finishing season two. 😃 | **<u>User 2</u>**<br>***Highest weight*: I'm losing my sight and going insane. Please tell me I'm ok.<br>***Lowest weight*: More university applications 😆 |
| **Non-depressed users** | |
| **<u>User 1</u>**<br>***Highest weight*: Did anyone watch the basketball match last night? It's classic and wonderful. Isn't it? 😮<br>***Lowest weight*: watching home alone 😛 It's such a classic movie. | **<u>User 2</u>**<br>***Highest weight*: Mom scores! Indeed, she will be bragging until tomorrow☺<br>***Lowest weight*: happy birthday to my son! |

# 4.3 Discussion

Sections *4.1* and *4.2* reported experiments on detecting depressed users from their textual content on microblog and social network platforms. As can be seen, MILA-SocNet and MIL-SocNet performed better than other replicated models in many dimensions. This section explains and discusses differences among MILA-SocNet, MIL-SocNet, and the previously published models.

As can be seen from *Figure 4-4* and *Figure 4-9*, MILA-SocNet, MIL-SocNet, and other previously published models can perform much better than a guessing line. This reflects that data from profile-based social media platforms can be used as a source of information to detect users with depression.

## 4.3.1 Comparisons with Replicated Models

We have compared the performance of our models with several previously published models. As can be seen from *Table 4-1* and *Table 4-6*, MILA-SocNet and MIL-SocNet outperformed all the other models in all dimensions. We further discuss several potential reasons for that. At the end of this section, we will explain what the limitations of the replicated models were.

While deep learning models can be trained on raw textual data, traditional machine learning models (e.g., the LIWC, language, topic, and Usr2Vec models) require feature extraction with external tools, which may introduce an additional risk of losing useful information from short textual data [132, 133]. For instance, misspelled and abbreviated words in messages may not be present in the dictionary of an extraction tool, resulting in words being mislabelled. A word “bad” may be misspelled as “bed” or “badddd”. The word may miss from a dictionary and be mislabelled. This results in the predictive performance of the model. This may be a reason why traditional machine learning techniques performed worse than our proposed models.

For example, a user has 5 posts following:

- I am feeling lonely.
- Tomorrow is weekends. Need to take a break!!!
- Cooking is my therapy.
- Tiredddd!!!
- Little things can make my big days.

These messages are extracted features using LIWC and then aggregated. This results in:

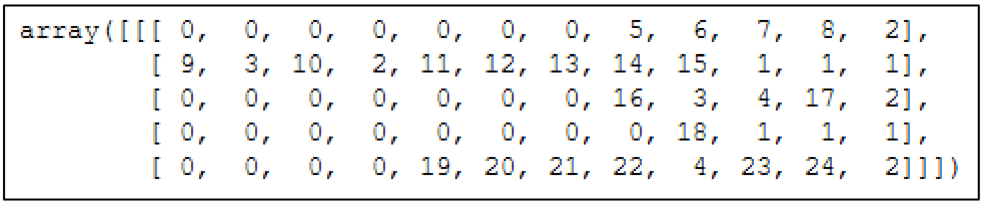
*Table 4-10 Feature extraction result using LIWC*

| **Post** | **Features** | | | |
|---|---|---|---|---|
| | **Word "I"** | **Positive words** | **Negative words** | **Health words** |
| 1 | 25 | 0 | 25 | 0 |
| 2 | 0 | 0 | 0 | 0 |
| 3 | 25 | 0 | 0 | 25 |
| 4 | 0 | 0 | 0 | 0 |
| 5 | 14.29 | 0 | 0 | 0 |
| **Aggregate** | **12.86** | **0** | **5** | **5** |

It can be seen that the post “Tiredddd!!!”, which LIWC cannot detect as a health-related category. When the message is changed to “Tired!!!”, LIWC can compute the health word category as 100. This can make a big difference for a model prediction.

Another reason for the performance gap may be that the sequential orderings of words in a post and messages posted on a timeline may have an effect on model performance. Training a predictive model with traditional machine learning methods requires aggregated data, which may cause the loss of contextual information compared with deep neural networks, which can learn from the raw textual data and sequential information in the data [134–137].

To train a LIWC predictive model, the aggregated result in Table 4-10 is used as a set of vectors for the user. In the case of training a deep learning model, the above messages are transformed to words to indexes represented in Figure 4-14. This raw data will be used for training a model. This illustrates that training a deep learning model can use all the raw and sequential data, while training a classical machine learning model requires feature extraction and data aggregation.



```
array([[[ 0,  0,  0,  0,  0,  0,  0,  5,  6,  7,  8,  2],
        [ 9,  3, 10,  2, 11, 12, 13, 14, 15,  1,  1,  1],
        [ 0,  0,  0,  0,  0,  0,  0, 16,  3,  4, 17,  2],
        [ 0,  0,  0,  0,  0,  0,  0,  0, 18,  1,  1,  1],
        [ 0,  0,  0,  0, 19, 20, 21, 22,  4, 23, 24,  2]]])
```

*Figure 4-14 Textual data transformed to word indexes*

Unlike the CNN model, MILA-SocNet and MIL-SocNet used bidirectional LSTMs that learns from forward and backward word sequences. Figure 4-15 depicts the learning process of bidirectional LSTMs. It can be seen that LSTMs can learn from the word's context of both. This means LSTMs is better when the meaning of text is depended on the entire sentence [138]. In contrast, CNN is better when learning from local regions (see Figure 4-16). LSTMs may have contributed to our proposed models outperforming CNN.

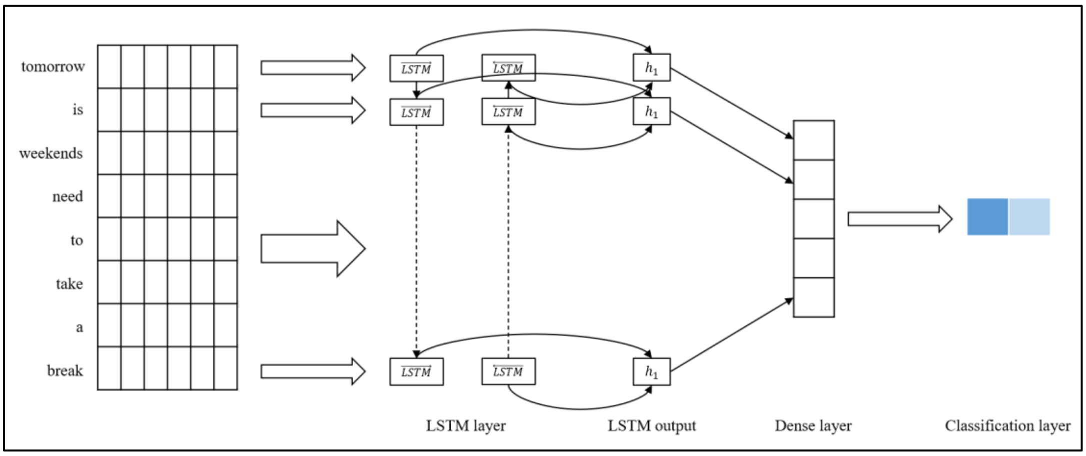


*Figure 4-15 LSTM model architecture with two channels for an example sentence*

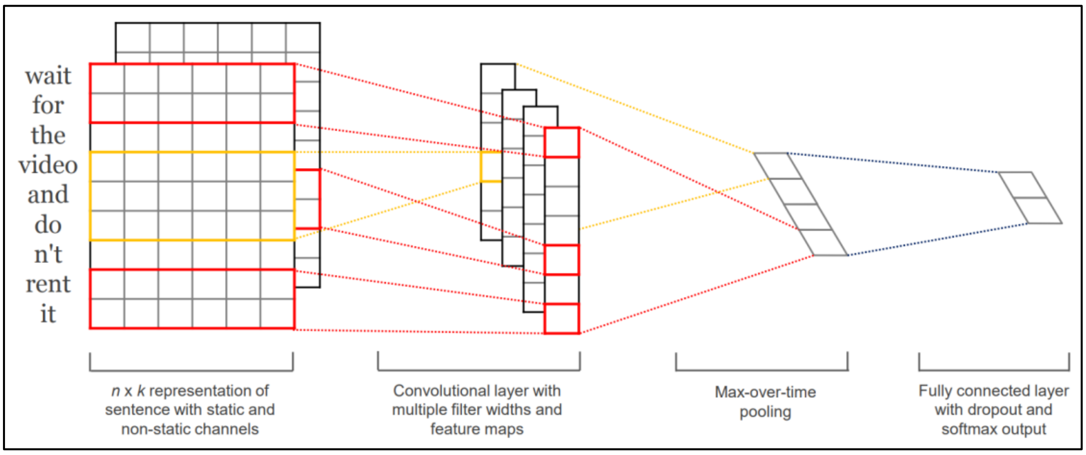

wait
for
the
video
and
do
n't
rent
it
n x k representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

*Figure 4-16 CNN model architecture with two channels for an example sentence (taken from* [139]*).*

Another different point between our proposed models and the deep learning model is that our models use attention mechanism that highlights words and posts relevant to mental health. This attention mechanism may help our proposed models surpassing the replicated deep learning model, even though our approach is also based on deep learning techniques.

The performance gaps between our proposed and the replicated models may come from the limitations of our experiments. For instance, when we replicated the LIWC model, it did not have the best performance as reported in the original study [24]. The paper reported that the best performing model not only used LIWC features but depression language and egocentric network measure features. Our study used only LIWC features to replicate the model, due to the limitation of feature acquisition, which may have had an impact on the performance of the LIWC model.

To train replicated models, misspelled and abbreviated words in our dataset were not corrected. This may also lower the performance of the replicated models, when comparing with our proposed models. It is worth noting that the lack of parameter tuning for these replicated models may also have an influence on the performance of the models and make our models outperforming. Our study used same machine learning techniques as reported in original studies. Using more complex machine learning algorithms and hyperparameters may improve the performance of the replicated models and make the models more competitive with our proposed models. This reflects that there is still room for improvement of the models.

### 4.3.2 Comparison Between MILA- and MIL-SocNet

Another important point to consider between MILA- and MIL-SocNet is whether the addition of anaphoric resolution encoder can improve the performance of the base MIL model. As can be seen in *Table 4-1* and *Table 4-6*, MILA-SocNet from microblog and social network datasets outperformed MIL-SocNet. This shows that the anaphora resolution encoding can improve the performance of predictive powers of MILA-SocNet. It can also identify posts related to the posters, rather than someone else. This is an important feature that has not been widely investigated in the field, and should be considered when designing future studies.

We further investigated whether there are any differences between the highest attention posts from MILA- and MIL-SocNet. We compared the highest attention posts from the two depressed users picked up in *Table 4-4*. From Table 4-11, it can clearly see that posts from MILA-SocNet show self-attentional focus related to mental issues, while MIL-SocNet focuses on topics not relevant to mental health issues and self-attention. This highlights that anaphora resolution could help to focus on self-disclosure of mental disorders.

*Table 4-11 Differences of highest attention posts from MILA- and MIL-SocNet*

| | **Depressed User 1** | **Depressed User 2** |
|---|---|---|
| *Highest attention posts from MILA-SocNet* | I was also dealing with depression and anxiety badly. School was hell. | I get some rest, take medication, and engage with what I like. These help me and I do not force myself to do things. |
| *Highest attention posts from MIL-SocNet* | Someone came up with a term that a community without an important person cannot provide somethings good for societies. | A mayor may govern a city but may only be in the position several years in a row and need to leave at least a few years before back. This seems to me that it's preventing newbies. |

### 4.3.3 Explainability Assessment

In this section, we further explored explainability by assessing the posts of two users from depressed and two from control groups randomly selected from those correctly labelled by MILA-SocNet. This will help us to understand insightful information of what users with and without depression were tweeting on their profiles.

This thematic assessment is followed and adapted approaches proposed by Makita et al. [140] and Berry et al. [141]. The former work presented a discourse coding framework used to screen tweets into 4 categories - mental health/illness as: (1) a medical condition; (2) a social issue; (3) a personal issue; and (4) information.

The latter explored why users tweeted mental health problems. They categorised tweets into 4 main themes: (1) sense of community; (2) raising awareness and combatting stigma; (3) safe space for expression; and (4) coping and empowerment. Additionally, 11 associated subthemes were investigated.

From their proposed coding frameworks, 120 tweets from two non-depressed and two depressed users achieved highest attention weights from MILA-SocNet were assessed.

Table 4-12 reports three domain disclosure of topics that users tweeted in their profiles. It can be seen that depressed users focused more on mental health disclosure, while non-depressed users did not publish a post related to mental health disclosure. Users with depression mentioned social issues such as "*…please don' t feel like you' re the only one. you are not alone*", and "*do things to make others feel good and it will come back to you*". We did not see any message about mental health/illness as information e.g., calls for participants, mental health professionals' shortage, and promotion of mental health resources, as reported in the original study.

Table 4-13 shows the number of self-focus and other focus associated with mental health and other topics. From the table, users with depression focused more on themselves and mental health topics, while users without depression tweets more on other topics and non-self-focus.

*Table 4-12 Disclosure topics*

| | **Depression** | **Control** |
|---|---|---|
| Mental health/illness as a medical condition | 15 | 0 |
| Mental health as a social issue | 7 | 0 |
| Other topics | 38 | 60 |
| **Total** | **60** | **60** |

*Table 4-13 Self-disclosure*

| | **Depression** | **Control** |
|---|---|---|
| **Mental health/illness topics** | | |
| • Self-focus | 22 | 0 |
| • Others | 0 | 0 |
| **Other topics** | | |
| • Self-focus | 29 | 3 |
| • Others | 9 | 57 |
| **Total** | **60** | **60** |

*Table 4-14 Themes and subthemes of why users tweeted about mental health*

| | **Depression** | **Control** |
|---|---|---|
| **Sense of community** | | |
| To send and receive messages of hope and support | 3 | 0 |
| To share and receive information | 5 | 0 |
| **Stigma and awareness** | | |
| To combat stigma | 9 | 0 |
| To raise awareness | 4 | 0 |
| To fight and campaign | 1 | 0 |
| **Total** | **22** | **0** |

The tweets associated with mental health disclosure were further categorised into themes and subthemes of why users tweeted about mental health.

Table 4-14 shows the number of tweets identified as each subtheme. The number of tweets from non-depressed users shows zero, because of no mental health disclosure. There is no report of using Twitter for escaping real-world, tracking back on their thoughts and feelings over time, and safe space for expression. Users with depression usually reported combating stigma and shared health information.

Our findings differ from the original studies, because they investigated URLs users tweeted for further information. Without performing that, we did not see many categories reported in the original study. According to Choudhury et al. [24], they found that users show high self-attentional focus, increased relational and medicinal concerns, similar to our finding.

### 4.3.4 Error Analysis

We further performed error analysis to observe mislabelled users from MILA-SocNet. This analysis was performed on the microblog dataset. There were 31 mislabelled users. 17 users without depression were incorrectly detected as depressed and 14 depressed users were wrongly labelled as non-depressed.

For incorrectly labelled users with depression, their top attention posts were related mental health issues. Some of them posted that they suffered from other mental health conditions. However, we used the search phrase “I was diagnosed with depression” to search target users with depression, which may contain users with other mental conditions. The original study proposing this data collection method also reported that the control users may be contaminated by users with various conditions [22]. We did not explore further to remove the control users reporting other health conditions from our dataset. This may be a reason why we labelled those users as non-depressed.

A user inaccurately detected without depression posted other languages. This issue reflects that our predictive model can be used for English-written posts only. Recently, the field of transfer learning for multi-language social media classification has emerged in the research community. Yang et al. proposed an effective and reusable election classifier for use on social media across multiple languages [142]. With some adaptations, this method can be applied to health classification from social media data, so that we can detect users with mental health problems from multi-language. Additionally, we may detect live events connected to people writing original language posts suffering from mental health problems [143].

Most of posts on five profiles contained retweets and replies. Additionally, four users posted many links on their microblogs. The users usually retweeted information or posted URLs they were interested, which our proposed model could not comprehend that information. The use of quote retweets or links may help better understanding additional information of what people talked about [144]. This approach may improve the performance of a predictive model and better comprehend users.

Finally, we explored predictive scores from our model and found that two users were received scores at the borderline between non-depression and depression. The profiles of two users were unexplainable. These users need the exploration for further information why the model could not correctly label those users for better understanding.

### 4.3.5 Model Limitations

Our models are nor free from limitations. Here we will explain some limitations to be addressed and improved in the future. *Our models could only detect depressed users with their openly expressed health*. The models may not work well on users who have not openly talked about their mental health issues. To detect users with mental health disorders, our models required a set of enough published posts to learn and make a prediction.

However, as shown above section, our model could correctly detect users with mental health conditions, even they did not publish the phrase “I was diagnosed with depression”. This shows that the architecture of our proposed model can be used to detect users with depression and its comorbidity [22]. In a future study, we plan to develop a predictive model trained with data from a target group publishing the phrase “I was diagnosed with depression” and test with data from another target group reporting other commodity or mental health conditions.

Our model was not trained with hyperparameters. However, we demonstrate the changes of parameters such the embedding dimensions, the length of word tokens, and the number of posts trained our model can improve the accuracy. The parameters of neural networks can also improve its performance. In this way, there is still room for improvement in the performance of our model.

The architecture of our proposed model focuses on the weights of posts relevant to mental health topics and associated with self-disclosure of posters. However, a post may consist of several sentences, which each of them may convey important information. The performance of our model would be improved by combining aspect-level sentiment analysis, which extracts information from each sentence [145].

Finally, as discussed in the section 2.6.4, there are many recent machine learning techniques and NLP not explored in the field of detecting users with mental health problems from their social media data. This is our future directions to improve the performance of predictive models.

## 4.4 Summary

This chapter presented results of applying MILA- and MIL-SocNet algorithms on profile-based social media datasets. The proposed models can work well on microblog and social network platforms. MILA-SocNet (with anaphora resolution) outperformed MIL-SocNet (without anaphora resolution). The former can instantly identify posts related to mental health issues and self-attentional focus. We compared the performance of our predictive models against a set of previous published models. The results show that our proposed models promisingly outperformed the replicated models in many dimensions.

# Chapter 5

# Ethical Concerns around the Use of Profile-Based Social Media Data for Mental Health Research

This chapter addresses the ethical concerns surrounding the use of profile-based social media data for mental health research. It begins with shaping key ethical concerns. A framework for using profile-based social media data for mental health research is then provided. Finally, we will explain factors affecting profiling for mental health and targeted advertisements.

## 5.1 Ethical Concerns

Although the current performance of predictive models is in the ascendant, reliable predictive models will eventually allow early detection and pave the way for health interventions in the forms of offering relevant health services or delivering useful health information links. By harnessing the capabilities offered to commercial entities on profile-based social media platforms, there is a potential to deliver real health benefits to users. However, ethical concerns have been raised about the use of publicly available data sources, profile-based social media data for health research, advertisements for mental health services, and social media platforms to report suicidal behaviour.

Several studies are particularly useful in highlighting the importance of ethical issues in this area of research. In a well-known example, researchers from Facebook and Cornell University [146] collected and used datasets from Facebook without offering the possibility to opt-out. This study was not approved by the Cornell University Institutional Review Board (IRB), "because this experiment was conducted by Facebook, Inc. for internal purposes, the Cornell University IRB determined that the project did not fall under Cornell's Human Research Protection Program" [147].

Also, another prominent study collected public Facebook posts and made the dataset publicly available to other researchers on the internet [148]. The posts were manually collected by accessing friend's profiles, and then they were anonymised. But even so, the posts could still be easily identified [149].

The above examples of collecting profile-based social media data from users and publicly available data on profile-based social media platforms raise concerns about ***privacy*** and ***informed consent***. This affects trust for social media users in transparency of research and ethics of researchers.

***General Data Protection Regulation (GDPR)*** helps to raise confidence in data safety and transparent analysis. However, *GDPR Article 9: Processing of special categories of personal data* specifically mentions that consent is not required if permission relates to personal data which are manifestly made public by the data subject. A core problem is the perception whether any data in public domain is automatically available for research. This is highly controversial from the ethical point of view, as the disruption presented by the wide availability of social network data impacts the norms that guide our perception on usage of our data for research. Ultimately, GDPR is focused on the process, not on the *objective* of the research, which is fundamental to shaping any research consent and the social consensus around it.

The ethical research practices of using profile-based social media data for research remain clearly undefined, incomplete, and inconsistent, particularly when working with information that is publicly available [150–152]. From those ethical concerns, we frame as a set of challenges to be overcome by different stakeholders. The next section will provide and explain a framework to deal with ethical considerations and practices when working with profile-based social media data for health analytics.

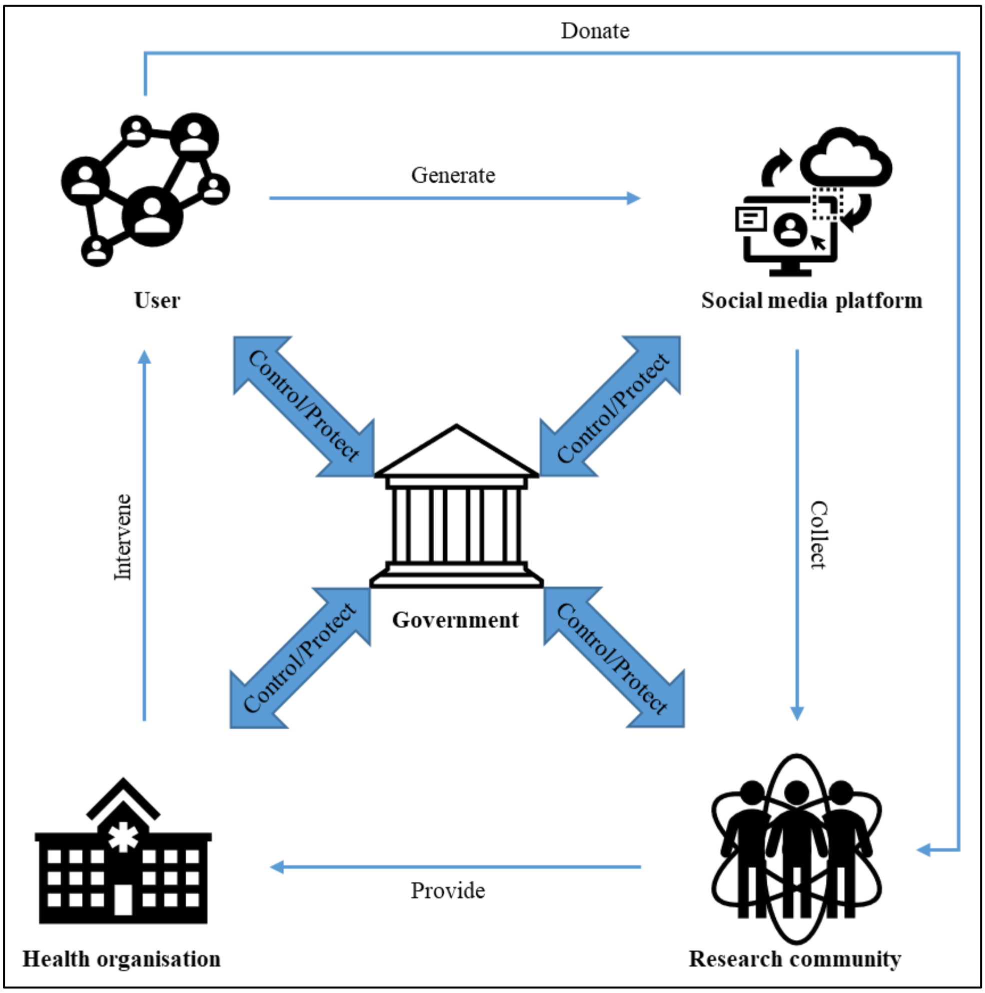


*Figure 5-1 A framework of key stakeholders associated with the use of profile-based social media platforms for health intervention*

## 5.2 A Framework for Using Social Media in Digital Health Interventions

Our framework for the use of profile-based social media platforms for health intervention involves five groups of stakeholders including users, social media platforms, research communities, health organisations, and governments. Figure 5-1 shows the framework and relevant stakeholders. We will provide details of each stakeholder below.

### 5.2.1 Users

This group includes all users currently using profile-based social media platforms. These people are considered as the most active and affected group, because they mainly are *providers* and *receivers* in this framework. In terms of providers, users generate data on the platforms and provide permissions to other relevant stakeholders to use their data for academic research and societal benefits. Receivers are clients who receive health services from stakeholders.

This group of people is aware that their data may be misused and analysed out of the original context of the research. For example, their data may be analysed for health issues affected the individual's insurance premiums [151] or explored by officers for prosecution [153].

In terms of informed consent for using their data, users have different ideas about their data privacy and can be divided into two groups: one for whom understanding the public nature of social media; and another for whom needing data ownership and private data. The former prefers not to receive informed consent but should be anonymised. The latter needs to be informed to use their generated data [150, 151, 153].

Users may also have concerns about the accuracy of the methods for data analysis. Due to concerns about data privacy, users are aware that they should not publish everything they think [153]. They also avoid posting when feeling uneasy or low [151]. Users believe that these issues may affect the performance of predictive models and analytics results [151, 153].

Sykora et al. conducted an online survey and a focus group. Three main ethical concerns from the participants were: (1) what information is being collected and processed: (2) what are the purposes of the analysis: and (3) whether users are informed about the data collection and analysis [150]. These concerns emphasise ethical practices for researchers to inform users before starting research from their data.

To overcome these issues, other stakeholders in the framework need to build trust for profile-based social media users to ensure not only that their data will be safe from misuse but also that the potential benefits outweigh any risks to privacy. This will make users feel comfortable and provide their data for research. They would also prefer to receive health analytics results and health services from stakeholders.

### 5.2.2 Social Media Platform

Profile-based social media is an online platform that allows users to publish posts, interact with other users, and communicate with others. The platform is not just an infrastructure but includes development, management, and policy teams. They need to maintain the stability of the infrastructures and create a good environment for users. The policy team should set a set of policies to support user’s trust and safeguard user’s privacy. For example, the company should prevent someone from access to sensitive data [154].

A well-known example highlighting this issue is Facebook data scandals. The company allowed researchers to access sensitive information and the data was subsequently used for 2016 US presidential election. Data was not used for an initial purpose of the research [155]. This made users lose trust of using their social media data for research [151]. Social media providers need to safeguard privacy of users and protect them from risks of data misuse.

Another concern for profile-based social media providers is publicly available data. This issue may undermine the trust and satisfaction of users to publish content on the platforms, because their data may be retrieved by researchers without their consent. However, it is difficult and challenging for researchers to receive informed consent from millions of users [156]. We suggest those platforms should develop a mechanism to restrict access to sensitive information and obtain informed consent from users before researchers can use user's data. Figure 5-2 depicts a mechanism for social media platforms to ask for research agreement from users. First, users register accounts on social media platforms, and then they accept with terms and conditions of using the platforms. The sites should develop an additional page to ask the users whether they agree to participate in research studies already approved by the platforms and IRBs. In this way, the users agree to provide their right to the platforms to screen forthcoming research and the platforms need to screen ethical and transparent research for their users.

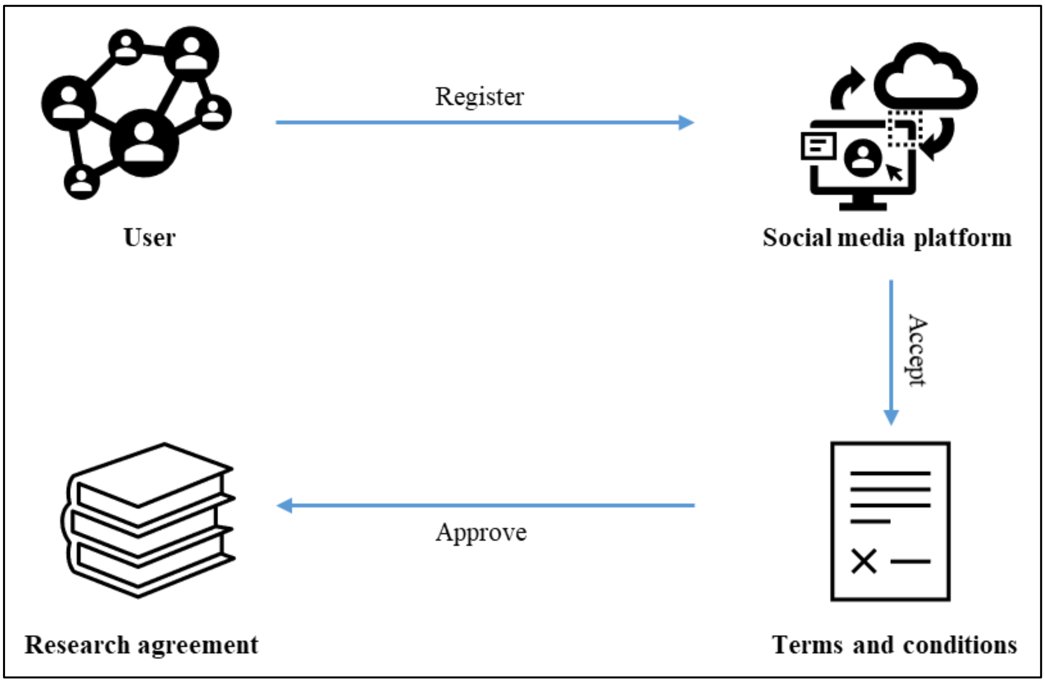


*Figure 5-2 A mechanism for social media platforms to ask for research agreement from users.*

However, the previous approach may make the users feel uneasy, because they do not know in which studies they are taking parts. The platforms may develop an additional function to notify users which studies are collecting your data. Consequently, the users can request research teams conducting the research to remove their data from the studies. This additional mechanism provides an option for the users to opt out in undesired studies. Due to the affiliation of researchers, users trust studies from some organisations such as universities [153].

These approaches would help researchers to access publicly available data ethically and at the same time it would allow users to feel comfortable to participate in research. This can increase public trust for users in transparent and ethical research and platforms can safeguard data privacy. The companies can also support societal benefits and development for public health research [157].

### 5.2.3 Researcher

Within the present framework, researchers are a group of academics from universities or people from non-profit research organisations who conduct studies on health analytics from profile-based social media data. Users trust not-for-profit researchers and academics rather than government and commercial organisations to use their profile-based social media data for research [153].

We surveyed a set of 82 studies analysing mental health from profile-based social media data, as displayed in *Figure 2-1*. We analysed those study methods of how they adopted a wide range of approaches to handle ethical constraints.

18% of them (15 studies) were approved by their IRBs [21, 27, 81, 83, 84, 95, 103, 34, 36, 38, 46, 60, 63, 64, 66]. 16% of them (13 studies) reported receiving informed consent from participants prior to data analysis [21, 30, 97, 100, 103, 34–36, 46, 60, 63, 64, 95]. For public data collected from crowdsourcing platforms, participants who opted in provided their consent to data sharing [21, 24, 63]. For myPersonality data, study [97] stated that the dataset itself had IRB approval, so the project did not report obtaining any further approval from the author’s institution. Study [158] also considered that no IRB approval was needed for using myPersonality data. Study [78] did not seek IRB approval, because of its reliance on Instagram data without personally identifiable information.

In studies [21, 43, 60, 63, 66, 75, 80, 82, 84, 85], the researchers reported the social network datasets collected from participants were anonymised. Study [84] removed names, user identifiers, and user identities, and the data collected had to be analysed after 3 months. All names were deleted from collected data before analysis

[84]. Tweet containing names or usernames were removed or replaced with other text [43, 75, 82]. Study [80] reported they removed user identifiers from tweets illustrated on the published paper.

Another concern for researchers is the accuracy and biases of methods of using social media data for health analytics. Users are aware of the analytics results which may not reflect faithfully their health conditions or public health [151, 153]. So, researchers need to prove the accuracy of predictive models and analytics results to increase trust of users to be profiled for health conditions.

It can be seen that those studies protected user's privacy by removing all identities and asking for informed consent to access profiles of users. These are ethical practices for researchers, as they can make participants feel comfortable to take parts in studies and gain trust from researchers.

### 5.2.4 Health Organisation

Another group of stakeholders is non-profit and licensed health organisations who intervene with people with certain health conditions and provide health services. Not-for-profit and licensed health organisations should be intermediary between social media platform providers and users, because users are feeling distrustful social media provides and commercial organisations [153]. They feel that academics are more ethical and not exploitative [153], while social media platforms have lost trust as data custodians and monetary profits [151], as can be seen from Facebook scandals [155].

Studies focused on the perspectives of participants in using social media for population health monitoring [159] and mental health disorders [151]. The authors and also the study [160] reported that most research participants agreed to have their public posts used for health monitoring, with anonymised data, although they also thought that informed consent would be necessary in some cases. Participants of these studies also believed that profiling for certain health conditions could help to increase access to health services, improve the diagnosis of health conditions, and target advertising for the provision of mental health care [151].

In this way, we suggest that non-profit and licensed health organisations should be representatives of other stakeholders to use health analytics results and target advertisements to users with certain health conditions. However, participants agreed that their privacy must outweigh societal benefits. If the organisations can reassure social media users that their privacy is safe and the top priority, this would make users feel more comfortable to receive health services and interventions [151].

### 5.2.5 Government

Governments and authorities are a group of people who create laws and are responsible for managing duties and services of stakeholders in the framework to use social media data for health analytics ethically and transparently. Governments and authorities should facilitate researchers, platform providers, and health organisation to generate public health, economic, scientific, and societal value for stakeholders. Authorities should also focus on the protection of unwanted surveillance and intrusion affecting to their citizens [161, 162]. They also need to penalise stakeholders who break the rules.

## 5.3 Ethical Considerations and Practices

Stakeholders relevant to using profile-based social media data for health analytics should be aware of ethical considerations and practices. All stakeholders should respect a code of conduct and any policy changes affecting to their roles.

### 5.3.1 Identity and User Protection

As explained in *section 5.2.3*, few studies have focused on ethical issues. Study [163] provided a taxonomy of ethical concepts to bear in mind when using Twitter data for public health studies. Studies [163, 164] reviewed and presented normative rules for using public Twitter data, including: paraphrasing collected posts, receiving informed consent from participants, hiding participant’s identity, and protecting collected data.

One possibility to reduce any conflict in this area of research is to anonymise the collected datasets to prevent the identification of users [165, 166]. In the US and UK, any research involving human participants must provide project information to IRBs or ethics committees for revision and obtain approval prior to data collection [167, 168].

Additionally, according to the Federal Policy for the Protection of Human Subjects ('Common Rule'), all studies conducted in the US are required to offer an opt-out for participants. However, private companies do not fall under this rule [147]. These ethical practices help to safeguard right of participant and their data privacy [156].

### 5.3.2 Policies and Terms of Service

Working with data collection from profile-based social media platforms, researchers need to take account of policies and terms of service (TOS). The study [169] reviewed TOS from over 100 different social media platforms. It shows that TOS from different platforms may state what types of data and in which way the data can be collected can be collect. TOS may be changed to respond to government regulations and some ethical concerns. This may have a big impact on using profile-based social media data for research. Researchers have to follow the terms of service of the social media platforms. When the policies are changed, it may have an effect on restrict to their services and sensitive information to protect data privacy.

For instance, the Facebook scandal and data misuse occurred in March 2018 [155]. This highly impacted on Facebook, which needed to change their policy and TOS [170]. Those changes restricted researchers' and developers' access to Facebook data. They needed to change their app mechanisms to comply with the terms and pass couple review processes (as elaborately explained in *appendix A.3.1*).

This may also prevent the researchers from accessing social network data and makes research progress delayed. For instance, with the TOS changes, all applications on the platform were required to re-submit for app review and pass business verification process. These further steps prevented some applications operating on Facebook, because some researchers of the applications could not provide business documents during the business verification. The data collection tool from Facebook elaborately explained in appendix B.2 could not be used for our earlier purposes due to this issue as well.

## 5.4 Success of Profiling Social Media for Health Conditions

In a recent survey on using social media data to identify users with depression, UK social network users expressed serious concerns about privacy risks and did not see the potential societal benefits outweighing these risks [151]. Thus, if these technologies are to have a meaningful impact on people's lives, increased importance must be placed on the transparency and trust of the analytics performed.

Achieving this trust is, to some extent, reinforced by the required compliance of any research with ethical codes and with our above framework, which would help to raise confidence in data safety and transparent analysis.

Another concern that would help profiling social media users for mental health to be success is online safe spaces. The users should collaboratively create online safe spaces and prevent online harms. The use of incivility, intolerance, and toxic language can prevent the users from being open-minded [171]. Users who posted about suffering from influenza, for instance, received sympathy, while users publishing about their infection with the human immunodeficiency virus (HIV) experienced stigma and discrimination from other users [162]. This means that users with certain medical conditions cannot openly express about their health issues. It would affect the effectiveness of health analytics from social media data because people cannot open-minded and disclose their conditions. With the issue around intolerance, this may introduce the problem of self-disclosure bias and performance of a predictive model, because users with mental problems may not want to publish their health conditions

To be successful, all stakeholders in our proposed framework need to facilitate each other and comply with ethical codes. Researchers, social media providers, and health organisations should collaborate together to provide better societal benefits for social media users. Governments and authorities need to protect their citizens by providing laws and regulations to control stakeholders to be ethical. All stakeholders need to ensure that societal benefits outweigh the risks to privacy of users.

Other third parties such as publishers should also monitor researchers and research studies to be ethical and transparent, when they are publishing articles by requiring their IRB approvals.

Finally, profiling for mental health conditions and target advertisements for mental health services can be achieved by collaborations from everyone. This is still a question to how we can control and convince everyone to protect privacy of and build trust for users. We all are as researchers and academics also need to be ethical and transparent and build trust for other stakeholders.

## 5.5 Discussion

These ethical and users concerns reported in this chapter are applied to our present study and linked up to earlier chapters. Compatible with the users concerns explained above, our study removes all important information and delete identities of users collected and used in our dataset. Without a mechanism to report users collected their data, we did not import them when collect their public data, as explained above. Social media platforms may provide a mechanism for researchers to import users, when we are collection their profiles, even publicly accessible, because some of them may want to publish their content for purposes but not for research.

Another concern by users and government is the transparency of model predictions and biases. This study demonstrates that we can build a predictive model with transparency by illustrating what the model focuses on and how it makes a decision.

Ethical practices are linked up to the papers included in our systematic review. This shows that many studies concerned user’s privacy by removing all identities and asking for informed consent to access profiles of users.

## 5.6 Summary

This chapter described the ethical concerns surrounding profiling for health analysis from profile-based social media data. We proposed a framework to tackle the issues of using publicly available data sources and targeted advertising for the provision of health care. The framework provides a detailed description of stakeholders and their responsibilities to facilitate the success of advertisements for mental health services. Overall, all the stakeholders need to build trust in those services and assure users that the potential benefits for them and for the society outweigh any risks to privacy.

# Chapter 6

# Conclusion

This chapter will summarise the research results and findings described throughout the thesis. It will highlight the implications of using machine learning on profile-based social media data sources for identifying users suffering from mental health disorders. Finally, this chapter also provides future directions for this area of research.

## 6.1 Research Objectives and Contributions

With the increasing number of people suffering from mental disorders and abundance of data produced on social media, monitoring users' mental health is becoming a promising area of research. This thesis developed predictive models for detecting depressed users from their profile-based social media data and identifying textual content associated with self-disclosure of mental health disorders. At the outset of this thesis, we listed four research objectives:

1. Survey the scope and limits of cutting-edge techniques for developing predictive models to identify profile-based social media users with mental health disorders.

2. Construct predictive models to classify profile-based social media users with mental health illnesses and identify textual content related to self-disclosure of mental illness.

3. Evaluate the feasibility of these models on examples of a microblogging platform and a social network platform.

4. Examine the impact of the work presented on ethical issues concerning the use of social media data for research.

We shall now discuss the findings of each one of these objectives.

### 6.1.1 Research Gaps and Growing Demands

The survey presented in *Chapter 2*, provided the scope and limits of the current state-of-the-art in mental health predictive analytics from social media.

The review showed a rise in popularity of this area, with more than half of the studies included focusing on depression. With the rising number of people suffering from depression, it is likely that the study of this disease will remain valuable in terms of societal benefits.

Articles published before 2017 intensively focused on the impact of classical NLP and manual feature extraction techniques such as LIWC on the accuracy of predictive models. The majority of predictive models developed from the reviewed studies used classical machine learning models such as regressions, SVMs, and decision trees. Statistical analysis was commonly used to understand the importance of individual features and their contribution to the model, typically requiring some mental health domain knowledge to interpret the results.

With deep learning techniques achieving notable performance in several fields, including medical image recognition, a promising research direction is to apply deep learning models with explainable and interpretable ability to classify users with mental illness from their profile-based social media data. Deep learning approaches have the ability to automatically extract features from raw text [172] and visual input data [173], so-called *feature learning* or *representation learning*.

This ability can help people outside the medical area to use these models. However, the challenge of deep learning is that it typically results in black box models whose inner workings are difficult understand due to the numbers, size, and nonlinearities of its layers [174, 175].

Explainability is introduced to eliminate the challenge. An explainable model is the capability of a machine learning model to present meaningful and insightful information to stakeholders to easily understand. Local Interpretable Model-agnostic Explanations (LIME) [176] and SHapley Additive exPlanations (SHAP) [177], for example, approaches are proposed to extract an individual prediction of a predictive model. These techniques work well in several types of data e.g., text and image and can be applied into classical machine learning e.g., random forests and neural networks. They also are the most comprehensive and dominant across many research areas [178].

Interpretability is the capability of a machine learning model to present a set of properties that the model bases a prediction on, e.g., *a linear regression model* can present the probability distribution of variables contributing a prediction result or *a decision tree* can provide internal nodes representing possible outcomes.

The difference between interpretability and explainability is that the former describes the set of inputs that a machine learning model uses to build its internal relations, while the latter focuses on giving the explanation about an individual prediction made by a model. Black-box, so-called explainable models, comparatively performs better than white-box, particularly so-called interpretable models, well in many tasks, but the challenge of the black-box is explainable predictions.

Only a minority of the studies surveyed in our systematic review used deep learning techniques, and they failed to provide explainable and interpretable results, but only focused on proposed machine learning architectures and model accuracy. Similarly, Linardatos reported that research studies focused on the performance of predictive models rather than the understanding behind these predictions. The field of eXplainable Artificial Intelligence (XAI) have attracted researchers after 2018 [178].

This highlights the importance of *explainability* and *interpretability* of machine learning in medical research, which is particularly relevant in mental health, as disease mechanisms are still not fully understood [179]. The extracted information from deep learning may provide us better understanding of how people cope with disease and help us diagnose sufferers better. LIME and SHAP are the example methods that can be applied to a predictive model to help us better understand how the model makes a decision. We also propose MIL- and MILA-SocNet that show how the model explains predictions.

### 6.1.2 Predictive Models for Detecting Users with Mental Disorders from Profile-Based Social Media Platforms

After identifying this research gap in the literature, we proposed deep learning predictive models to classify social media users with mental health illness and provide explainable and interpretable power. A social media dataset contains labels not only for users but for individual posts as well, as explained in *Section 3.2*. Considering the nature of a dataset, an algorithm needs to be optimised for a particular type of a dataset [180–182]. We found the paradigm of multiple instance learning (MIL) to be well-suited to a profile-based social media dataset.

The two models developed, with and without anaphora resolution, were evaluated against a set of previously published models ranging from classical machine learning to deep learning techniques in *Chapter 4*. The experiments found that our algorithms (MIL-SocNet and MILA-SocNet) are suitable for a profile-based social media dataset and a task of classifying social media users with mental disorders, as can be seen from promising and comparable results against other published algorithms.

### 6.1.3 Applicability of Multiple Instance Learning Algorithms

Our proposed models were trained on a microblogging platform dataset from Twitter and a social network dataset from Facebook. We found that both models were effective in identifying markers of mental illness, which were confirmed by our subjects found to be experiencing a range of emotions, e.g., feeling broken-hearted, suffering a loss, experiencing school problems, as shown in *Table 4-4* and *Table 4-9*. These all match existing knowledge of depression, however, in the future this may pave the way for detecting new and emerging features associated with mental disorders, such as long-COVID, furlough concerns, etc.

These results confirmed our hypothesis that MIL is well-suited to profile-based social media datasets and can identify markers associated with mental disorders. As can be seen, MIL itself can detect and label posts relevant to mental health topics, while other machine learning approaches required additional methods, such as statistical analysis, to extract usable knowledge.

### 6.1.4 Ethical Standards for the Study of Mental Disorders from Profile-Based Social Media Platforms

The issue for user consent is particularly relevant when using data posted onto social media, so we surveyed the main issues that users express concern about with regards to the use of their data.

One clear concern is the risk to users' privacy, with social media providers not being trusted as data custodians, as highlighted in several studies [151, 153, 183]. So, one clear direction for the success of profiling social media for mental health is to build trust for the users. We must assure users that their data will not be misused and accessed by other third parties.

Additionally, we surveyed studies on ethical issues concerning the use of social media data for profiling mental health. We summarised and proposed a framework for the key stakeholders including users, social media providers, researchers, health organisations, and governments. All these stakeholders need to build trust for the users by ensuring the security of data collection, removing identifiable information, providing effective opt-out options, using data transparently and with clear goals with direct mapping to benefits for individual users and/or society.

Additionally, all stakeholders from our purposed framework should collaborate to foster societal benefits from profiling social media data for health users. This also comes to the question whether social media platforms should help research societies and allow researchers to access their data for academic purposes easier. For example, Facebook should permit research accessing its platform to foster health research from its data. Noticing, relatively few studies used Facebook data to profile users' health and mental problems. It also highlights that the researchers should make decision ethically for data collection from social media platforms by considering the context of their research studies and TOS of social media platforms [169].

Engagement with the users themselves is needed to create an environment in which users are comfortable to freely and openly express their feelings, thoughts, and expressions about their mental health issues, and where interaction with other users can provide them with valuable helps and interventions.

Finally, GDPR regulation requires the transparency of computing algorithms. A model or algorithm needs to provide "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject" [184, 185]. Our work proves that our proposed models can provide insightful information of how the model makes decisions as shown in *sections 4.1.5.2* and *4.2.5.2*.

## 6.2 Future Directions

As this research showed, it is feasible to use machine learning for identifying profile-based social media users with mental health problems. This section will highlight important open questions which have the potential to further develop the field and propose a framework for delivering digital interventions.

***The combination of different feature spaces and data sources****:* The experiments presented relied solely on textual data for modelling. The combinations of different feature spaces such as images, links, and comments posted on profile-based social media can improve the performance of a model [60]. It is also possible to link profile-based social media data to data from other sources such as electronic health records or wearables and other smart devices to provide a richer data space for detecting mental and physical health problems [186].

On acceptance of profiling profile-based social media for mental disorders from users, supporting methods can be developed to translate this innovation into practice and provide digital help to individuals. Figure 6-1 depicts the potential ecosystem of applications for users and mental health services. The applications include:

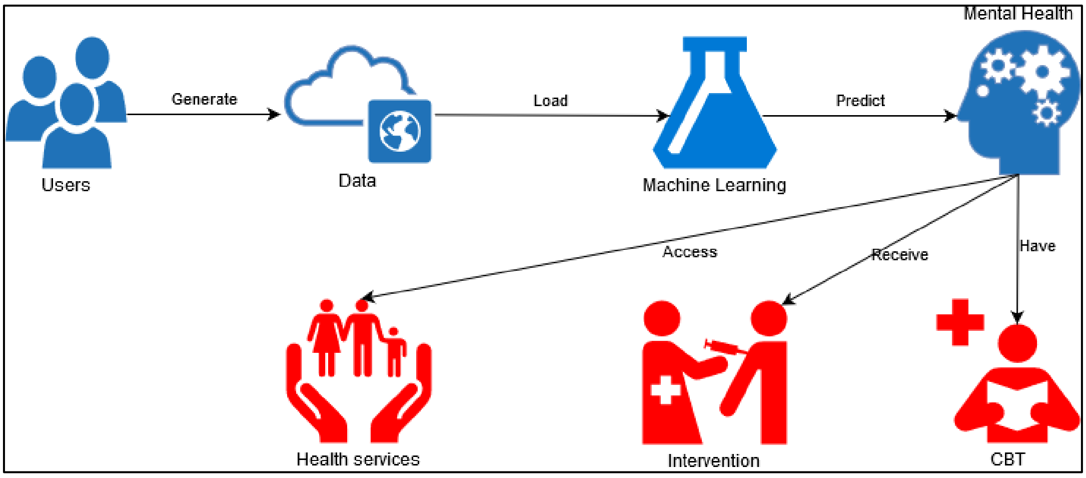
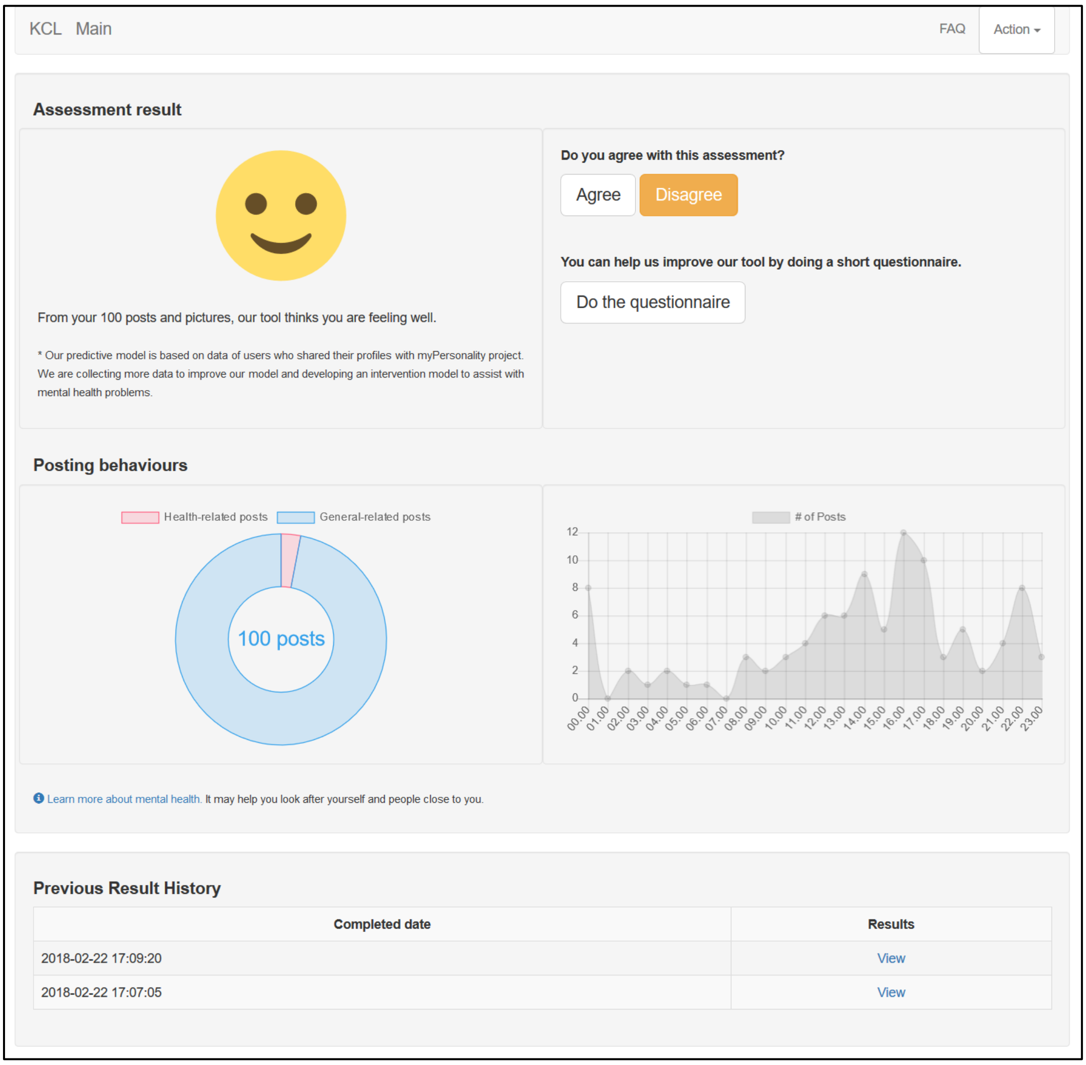


*Figure 6-1 Conceptual view of social network-based mental health research.*

***Digital intervention****:* A reliable predictive model will eventually allow early detection and pave the way for health interventions in the forms of promoting relevant health services or delivering useful health information links. By harnessing the capabilities offered to commercial and non-commercial entities on profile-based social media, there is a potential to deliver real health benefits to users such as online mental health services to a person needing mental health support.

***Digital intervention tools*** can be used to provide interventions for users identifying as suffering from mental health issues. Having been identifying by the tool, the user would be presented with a predictive result, and based on the result, presented the findings together with links to appropriate mental health services. Example screens are shown in Figure 6-2 and Figure 6-3. Such tools could play a significant role in getting the patients into services sooner, when their treatment cost is lower and outcomes better – a major challenge for mental health service providers.



**Assessment result**

From your 100 posts and pictures, our tool thinks you are feeling well.

* Our predictive model is based on data of users who shared their profiles with myPersonality project. We are collecting more data to improve our model and developing an intervention model to assist with mental health problems.

**Do you agree with this assessment?**

**You can help us improve our tool by doing a short questionnaire.**

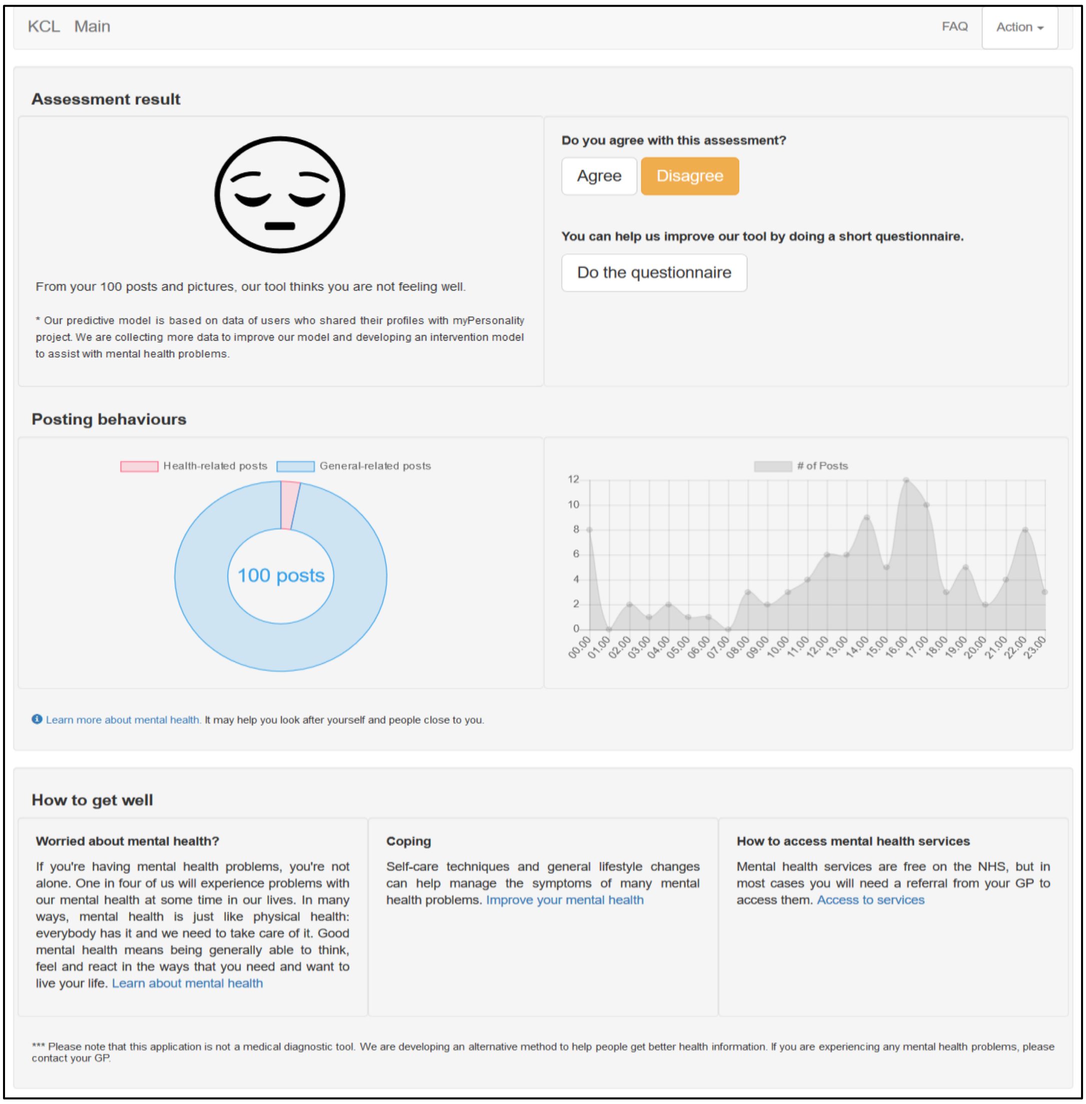
**Posting behaviours**

Learn more about mental health. It may help you look after yourself and people close to you.

**Previous Result History**

| Completed date | Results |
|---|---|
| 2018-02-22 17:09:20 | View |
| 2018-02-22 17:07:05 | View |

*Figure 6-2 The example of digital intervention tool shows the screen when not detecting users with mental disorders*



KCL Main FAQ Action ▾

**Assessment result**

From your 100 posts and pictures, our tool thinks you are not feeling well.

* Our predictive model is based on data of users who shared their profiles with myPersonality project. We are collecting more data to improve our model and developing an intervention model to assist with mental health problems.

**Do you agree with this assessment?**

Agree Disagree

**You can help us improve our tool by doing a short questionnaire.**

Do the questionnaire

**Posting behaviours**

Learn more about mental health. It may help you look after yourself and people close to you.

**How to get well**

**Worried about mental health?**

If you're having mental health problems, you're not alone. One in four of us will experience problems with our mental health at some time in our lives. In many ways, mental health is just like physical health: everybody has it and we need to take care of it. Good mental health means being generally able to think, feel and react in the ways that you need and want to live your life. Learn about mental health

**Coping**

Self-care techniques and general lifestyle changes can help manage the symptoms of many mental health problems. Improve your mental health

**How to access mental health services**

Mental health services are free on the NHS, but in most cases you will need a referral from your GP to access them. Access to services

*** Please note that this application is not a medical diagnostic tool. We are developing an alternative method to help people get better health information. If you are experiencing any mental health problems, please contact your GP.

*Figure 6-3 The example of digital intervention tool shows the screen when detecting users with mental disorders.*

***Medical decision support system:*** Profile-based social media data can offer many insights about the health of users. A predictive model with explainability and interpretability power can help doctors to receive more useful information to make a treatment decision. Symptoms of mental disorders may be extracted from posts of users as shown in *Table 4-4* and *Table 4-9*. The doctors can use that information to diagnose patients.

## 6.3 Closing Remarks

The potential impact of this research lies in its ability to offer depressed social media users suitable help at an early stage. Depression, as well as other mental health disorders, is easier and cheaper to treat the sooner patients seek help, and offering these services to people may significantly improve their outcomes. Targeted advertising by mental health charities may be seen as intrusive but is essentially no different than companies advertising their products to potential consumers based on their web activity.

Early research into the public perception of this type of data usage shows that there is some public scepticism about this approach [150, 151]. To overcome this animosity towards using social media data for mental health prediction modelling, we believe that future research in the area should focus on *explainability* and *interpretability*. We have shown that deep learning models perform well, but they offer no explanation of their decision-making process [187, 188]. Extraction of patterns from the models can improve interpretability, as we demonstrated with the social media post weight examples, and systematic sampling should be employed to achieve acceptable levels of trust. To gauge how acceptable these techniques are to the public, we intend to work with citizen juries to explore whether an improvement in explainability can lead to a change in public opinion [189].

We are now closing with take away messages as follow:

- Although the current performance of predictive models for detecting profile-based social media users with mental disorders is in the ascendant, reliable predictive models will eventually allow early detection and pave the way for health interventions and medical decision support system.
- A new model in this field needs to come up with *explainability* and *interpretability* to provide experts and practitioners a remarkable insight about the symptom of mental disorders.
- Finally, researchers, health organisations, social media providers, and governments, including users themselves must not only prioritise individual and societal benefits but also eliminate risks to privacy to put forward the success of the profiling social media for mental health and digital interventions.

# References

1 World Health Organization, Depression, 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression.

2 Public Health England, Health matters: reducing health inequalities in mental illness, 2018. [Online]. Available: https://www.gov.uk/government/publications/health-matters-reducing-health-inequalities-in-mental-illness/health-matters-reducing-health-inequalities-in-mental-illness.

3 D. E. Bloom, E. Cafiero, E. Jané-Llopis, S. Abrahams-Gessel, R. L. Bloom, S. Fathima, B. A. Feigl, T. Gaziano, M. Mowafi, A. Pandya, K. Prettner, L. Rosenberg, B. Seligman, A. Z. Stein, and C. Weinstein, The Global Economic Burden of Noncommunicable Diseases, *World Economic Forum*, 2011. [Online]. Available: http://www3.weforum.org/docs/WEF_Harvard_HE_GlobalEconomicBurden NonCommunicableDiseases_2011.pdf: . Webcite: 6CSThUnbF.

4 World Health Organisation, Depression, 2018. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression.

5 WHO, Depression, a global public health concern, *WHO Department of Mental Health and Substance Abuse*, 2012. [Online]. Available: http://www.who.int/mental_health/management/depression/who_paper_depression_wfmh_2012.pdf: . Webcite: 6j4f0tPeN.

6 WHO, Mental disorders, 2016. [Online]. Available:

http://www.who.int/mediacentre/factsheets/fs396/en/: . Webcite: 6oiTFbsMZ. [Accessed: 04-Mar-2017].

7 D. Chisholm, K. Sweeny, P. Sheehan, B. Rasmussen, F. Smit, P. Cuijpers, and S. Saxena, Scaling-up treatment of depression and anxiety: a global return on investment analysis, *The Lancet Psychiatry*, vol. 3, no. 5, pp. 415–424, May 2016. doi:10.1016/S2215-0366(16)30024-4.

8 P. S. Wang, S. Aguilar-Gaxiola, J. Alonso, M. C. Angermeyer, G. Borges, E. J. Bromet, R. Bruffaerts, G. de Girolamo, R. de Graaf, O. Gureje, J. M. Haro, E. G. Karam, R. C. Kessler, V. Kovess, M. C. Lane, S. Lee, D. Levinson, Y. Ono, M. Petukhova, J. Posada-Villa, S. Seedat, and J. E. Wells, Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys, *Lancet*, vol. 370, no. 9590, pp. 841–850, Sep. 2007. doi:10.1016/S0140-6736(07)61414-7.

9 L. Radloff, The CES-D Scale: A self-report depression scale for use in general populations, *Appl. Psychol. Meas.*, vol. 1, pp. 385–400, 1977.

10 M. Hamilton, Development of a rating scale for primary depressive illness., *Br. J. Soc. Clin. Psychol.*, vol. 6, no. 4, pp. 278–296, 1967. doi:10.1111/j.2044-8260.1967.tb00530.x. Medline:6080235.

11 W. W. K. ZUNG, A Self-Rating Depression Scale, *Arch. Gen. Psychiatry*, vol. 12, no. 1, p. 63, 1965. doi:10.1001/archpsyc.1965.01720310065008.

12 Internet World Stats, Internet usage statistics: World Internet Users and 2020 Population Stats, 2020. [Online]. Available: https://www.internetworldstats.com/stats.htm.

13 Facebook Inc, Facebook Reports First Quarter 2021 Results, 2021. [Online]. Available: https://investor.fb.com/investor-news/press-release-details/2021/Facebook-Reports-First-Quarter-2021-Results/default.aspx.

14 Twitter Inc, Q1 2021 Letter to Shareholders, 2021. [Online]. Available: https://s22.q4cdn.com/826641620/files/doc_financials/2021/q1/Q1’21-Shareholder-Letter.pdf.

15 I. Korkontzelos, A. Nikfarjam, M. Shardlow, A. Sarker, S. Ananiadou, and G. H. Gonzalez, Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts, *J. Biomed. Inform.*, vol. 62, pp. 148–158, Aug. 2016. doi:10.1016/j.jbi.2016.06.007.

16 J. Ive, G. Gkotsis, R. Dutta, R. Stewart, and S. Velupillai, Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health, in *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 2018, pp. 69–77. doi:10.18653/v1/W18-0607.

17 G. Gkotsis, A. Oellrich, S. Velupillai, M. Liakata, T. J. P. Hubbard, R. J. B. Dobson, and R. Dutta, Characterisation of mental health conditions in social media using Informed Deep Learning, *Sci. Rep.*, vol. 7, p. 45141, Mar. 2017. doi:10.1038/srep45141.

18 J. H. Lee, S. H. Jung, and J. Park, The role of entropy of review text sentiments on online WOM and movie box office sales, *Electron. Commer. Res. Appl.*, vol. 22, pp. 42–52, Mar. 2017. doi:10.1016/j.elerap.2017.03.001.

19 J. Wilkerson and A. Casas, Large-Scale Computerized Text Analysis in

Political Science: Opportunities and Challenges, *Annu. Rev. Polit. Sci.*, vol. 20, no. 1, pp. 529–544, May 2017. doi:10.1146/annurev-polisci-052615-025542.

20 A. Wongkoblap, M. A. Vadillo, and V. Curcin, Researching Mental Health Disorders in the Era of Social Media: Systematic Review., *J. Med. Internet Res.*, vol. 19, no. 6, p. e228, Jun. 2017. doi:10.2196/jmir.7215. Medline:28663166.

21 A. G. Reece and C. M. Danforth, Instagram photos reveal predictive markers of depression, *EPJ Data Sci.*, vol. 6, no. 1, p. 15, Dec. 2017. doi:10.1140/epjds/s13688-017-0110-z.

22 G. Coppersmith, M. Dredze, and C. Harman, Quantifying Mental Health Signals in Twitter, in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014, pp. 51–60. doi:10.3115/v1/W14-3207.

23 A. H. Yazdavar, H. S. Al-Olimat, M. Ebrahimi, G. Bajaj, T. Banerjee, K. Thirunarayan, J. Pathak, and A. Sheth, Semi-Supervised Approach to Monitoring Clinical Depressive Symptoms in Social Media, in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 - ASONAM '17*, 2017, pp. 1191–1198. doi:10.1145/3110025.3123028.

24 M. De Choudhury and M. Gamon, Predicting Depression via Social Media, *Proc. Seventh Int. AAAI Conf. Weblogs Soc. Media*, vol. 2, pp. 128–137, 2013.

25 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, *J.*

*Med. Internet Res.*, vol. 21, no. 6, p. e14199, Jun. 2019. doi:10.2196/14199.

26 M. De Choudhury, S. Counts, and E. Horvitz, Predicting postpartum changes in emotion and behavior via social media, *Proc. ACM Annu. Conf. Hum. Factors Comput. Syst.*, pp. 3267–3276, 2013. doi:10.1145/2470654.2466447.

27 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. doi:10.3115/v1/W15-1201.

28 J. C. Eichstaedt, R. J. Smith, R. M. Merchant, L. H. Ungar, P. Crutchley, and D. Preot, Facebook language predicts depression in medical records, pp. 2–7, 2018. doi:10.1073/pnas.1802331115.

29 H. A. Schwartz, J. Eichstaedt, M. L. Kern, G. Park, M. Sap, D. Stillwell, M. Kosinski, and L. Ungar, Towards Assessing Changes in Degree of Depression through Facebook, *Proc. Work. Comput. Linguist. Clin. Psychol. From Linguist. Signal to Clin. Real.*, pp. 118–125, 2014.

30 M. De Choudhury, S. Counts, E. J. Horvitz, and A. Hoff, Characterizing and predicting postpartum depression from shared facebook data, in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*, 2014, pp. 626–638. doi:10.1145/2531602.2531675.

31 S. Angelidis and M. Lapata, Multiple Instance Learning Networks for Fine-Grained Sentiment Analysis, *Trans. Assoc. Comput. Linguist.*, vol. 6, pp. 17–31, 2018.

32 B. Aktaş, T. Scheffler, and M. Stede, Anaphora Resolution for Twitter Conversations: An Exploratory Study, in *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, 2018, pp. 1–10. doi:10.18653/v1/W18-0701.

33 K. Kang, C. Yoon, and E. Y. Kim, Identifying depressive users in Twitter using multimodal analysis, in *2016 International Conference on Big Data and Smart Computing (BigComp)*, 2016, pp. 231–238. doi:10.1109/BIGCOMP.2016.7425918.

34 S. Park, I. Kim, S. W. Lee, J. Yoo, B. Jeong, and M. Cha, Manifestation of Depression and Loneliness on Social Networks, in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, 2015, vol. 20, no. 6, pp. 557–570. doi:10.1145/2675133.2675139.

35 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)*, 2015, pp. 361–364. doi:10.1109/WI-IAT.2015.166.

36 S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki, Recognizing Depression from Twitter Activity, *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst. - CHI '15*, pp. 3187–3196, 2015. doi:10.1145/2702123.2702280.

37 D. Preoiuc-Pietro, M. Sap, H. A. Schwartz, and L. Ungar, Mental Illness Detection at the World Well-Being Project for the CLPsych 2015 Shared Task,

in *the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015, pp. 40–45.

38 M. Mitchell, K. Hollingshead, and G. Coppersmith, Quantifying the Language of Schizophrenia in Social Media, in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015, pp. 11–20.

39 D. Preoţiuc-Pietro, J. Eichstaedt, G. Park, M. Sap, L. Smith, V. Tobolsky, H. A. Schwartz, and L. Ungar, The role of personality, age, and gender in tweeting about mental illness, in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015, pp. 21–30. doi:10.3115/v1/W15-1203.

40 T. Pedersen, Screening Twitter Users for Depression and PTSD with Lexical Decision Lists, in *the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015, pp. 46–53. doi:10.3115/v1/w15-1206.

41 P. Resnik, W. Armstrong, L. Claudino, T. Nguyen, V. 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*, 2015, vol. 1, no. 2014, pp. 99–107.

42 P. Resnik, W. Armstrong, L. Claudino, T. Nguyen, V.-A. Nguyen, and J. Boyd-Graber, The University of Maryland CLPsych 2015 Shared Task System, in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical*

*Psychology: From Linguistic Signal to Clinical Reality*, 2015, no. c, pp. 54–60.

43 M. L. Wilson, S. Ali, and M. F. Valstar, Finding information about mental health in microblogging platforms, *Proc. 5th Inf. Interact. Context Symp. - IIiX '14*, pp. 8–17, 2014. doi:10.1145/2637002.2637006.

44 V. M. Prieto, S. Matos, M. Álvarez, F. Cacheda, and J. L. Oliveira, Twitter: a good place to detect health conditions, *PLoS One*, vol. 9, no. 1, p. e86191, Mar. 2014. doi:10.1371/journal.pone.0086191. Medline:24489699.

45 H. Lin, J. Jia, Q. Guo, Y. Xue, Q. Li, J. Huang, L. Cai, and L. Feng, User-level psychological stress detection from social media using deep neural network, *Proc. ACM Int. Conf. Multimed. - MM '14*, pp. 507–516, 2014. doi:10.1145/2647868.2654945.

46 S. Park, S. W. Lee, J. Kwak, M. Cha, and B. Jeong, Activities on Facebook reveal the depressive state of users, *J. Med. Internet Res.*, vol. 15, no. 10, p. e217, Oct. 2013. doi:10.2196/jmir.2718. Medline:24084314.

47 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 *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7867 LNAI, 2013, pp. 201–213. doi:10.1007/978-3-642-40319-4_18.

48 X. Wang, C. Zhang, and L. Sun, An Improved Model for Depression Detection in Micro-Blog Social Network, *2013 IEEE 13th Int. Conf. Data Min. Work.*, pp. 80–87, 2013. doi:10.1109/ICDMW.2013.132.

49 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 *2013 IEEE Virtual Reality (VR)*, 2013, pp. 1–4. doi:10.1109/VR.2013.6549431.

50 M. De Choudhury, S. Counts, and E. Horvitz, Social media as a measurement tool of depression in populations, in *Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13*, 2013, pp. 47–56. doi:10.1145/2464464.2464480.

51 X. Yang, R. McEwen, L. R. Ong, and M. Zihayat, A big data analytics framework for detecting user-level depression from social networks, *Int. J. Inf. Manage.*, vol. 54, p. 102141, 2020. doi:https://doi.org/10.1016/j.ijinfomgt.2020.102141.

52 M. Stankevich, I. Smirnov, N. Kiselnikova, and A. Ushakova, Depression Detection from Social Media Profiles, in *Data Analytics and Management in Data Intensive Domains*, 2020, pp. 181–194.

53 M. Stankevich, A. Latyshev, E. Kuminskaya, I. Smirnov, and O. Grigoriev, Depression detection from social media texts, in *Data Analytics and Management in Data Intensive Domains: XXI In-ternational Conference DAMDID/RCDL'2019 (October 15--18, 2019, Kazan, Russia): Conference Proceedings. Edited by Alexander Elizarov, Boris Novikov, Sergey Stupnikov.--Kazan: Kazan Federal Uni*, 2019, p. 352.

54 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 *Proceedings of the Twenty-Sixth International Joint*

*Conference on Artificial Intelligence*, 2017, pp. 3838–3844. doi:10.24963/ijcai.2017/536.

55 S. Almouzini, M. Khemakhem, and A. Alageel, Detecting Arabic Depressed Users from Twitter Data, *Procedia Comput. Sci.*, vol. 163, pp. 257–265, 2019. doi:10.1016/j.procs.2019.12.107.

56 Y. Huang, C.-F. Chiang, and A. Chen, Predicting Depression Tendency based on Image, Text and Behavior Data from Instagram, in *Proceedings of the 8th International Conference on Data Science, Technology and Applications*, 2019, pp. 32–40. doi:10.5220/0007833600320040.

57 S. Maxim, N. Ignatiev, and I. Smirnov, Predicting Depression with Social Media Images, in *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods*, 2020, pp. 235–240. doi:10.5220/0009168602350240.

58 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. doi:10.1145/3372278.3391932.

59 X. Chen, M. D. Sykora, T. W. Jackson, and S. Elayan, What about Mood Swings: Identifying Depression on Twitter with Temporal Measures of Emotions, in *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, 2018, pp. 1653–1660. doi:10.1145/3184558.3191624.

60 S. Chandra Guntuku, D. Preotiuc-Pietro, J. C. Eichstaedt, and L. H. Ungar,

What Twitter Profile and Posted Images Reveal about Depression and Anxiety, *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 13, no. 01, pp. 236–246, 2019.

61 S. Amir, G. Coppersmith, P. Carvalho, M. J. Silva, and B. C. Wallace, Quantifying Mental Health from Social Media with Neural User Embeddings, in *Proceedings of the 2nd Machine Learning for Healthcare Conference*, 2017, vol. 68, pp. 306–321.

62 A. Husseini Orabi, P. Buddhitha, M. Husseini 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. doi:10.18653/v1/W18-0609.

63 A. G. Reece, A. J. Reagan, K. L. M. M. Lix, P. S. Dodds, C. M. Danforth, and E. J. Langer, Forecasting the onset and course of mental illness with Twitter data, *Sci. Rep.*, vol. 7, no. 1, p. 13006, Dec. 2017. doi:10.1038/s41598-017-12961-9. Medline:29021528.

64 Q. Cheng, T. M. Li, C.-L. L. Kwok, T. Zhu, and P. S. Yip, Assessing Suicide Risk and Emotional Distress in Chinese Social Media: A Text Mining and Machine Learning Study, *J. Med. Internet Res.*, vol. 19, no. 7, p. e243, Jul. 2017. doi:10.2196/jmir.7276. Medline:28694239.

65 A. Benton, M. Mitchell, and D. Hovy, Multitask Learning for Mental Health Conditions with Limited Social Media Data, in *Proceedings of the 15th Conference of the {E}uropean Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, vol. 1, pp. 152–162. doi:10.18653/v1/e17-1015.

66 A. H. Yazdavar, M. S. Mahdavinejad, G. Bajaj, W. Romine, A. Sheth, A. H. Monadjemi, K. Thirunarayan, J. M. Meddar, A. Myers, J. Pathak, and P. Hitzler, Multimodal mental health analysis in social media, *PLoS One*, vol. 15, no. 4, p. e0226248, Apr. 2020. doi:10.1371/journal.pone.0226248.

67 M. Y. Wu, C.-Y. Shen, E. T. Wang, and A. L. P. Chen, A deep architecture for depression detection using posting, behavior, and living environment data, *J. Intell. Inf. Syst.*, vol. 54, no. 2, pp. 225–244, Apr. 2020. doi:10.1007/s10844-018-0533-4.

68 C. Y. Chiu, H. Y. Lane, J. L. Koh, and A. L. P. Chen, Multimodal depression detection on instagram considering time interval of posts, *J. Intell. Inf. Syst.*, May 2020. doi:10.1007/s10844-020-00599-5.

69 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 Med. Informatics*, vol. 8, no. 7, p. e17958, Jul. 2020. doi:10.2196/17958.

70 P. Mann, A. Paes, and E. H. Matsushima, See and Read: Detecting Depression Symptoms in Higher Education Students Using Multimodal Social Media Data, *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 14, no. 1, pp. 440–451, May 2020.

71 G. Li, B. Li, L. Huang, and S. Hou, Automatic Construction of a Depression-Domain Lexicon Based on Microblogs: Text Mining Study, *JMIR Med. Informatics*, vol. 8, no. 6, p. e17650, Jun. 2020. doi:10.2196/17650.

72 H. S. ALSAGRI and M. YKHLEF, Machine Learning-Based Approach for Depression Detection in Twitter Using Content and Activity Features, *IEICE*

*Trans. Inf. Syst.*, vol. E103.D, no. 8, pp. 1825–1832, Aug. 2020. doi:10.1587/transinf.2020EDP7023.

73 T. Gui, Q. Zhang, L. Zhu, X. Zhou, M. Peng, and X. Huang, Depression Detection on Social Media with Reinforcement Learning, 2019, pp. 613–624. doi:10.1007/978-3-030-32381-3_49.

74 X. Chen, M. Sykora, T. Jackson, S. Elayan, and F. Munir, Tweeting Your Mental Health: an Exploration of Different Classifiers and Features with Emotional Signals in Identifying Mental Health Conditions, 2018. doi:10.24251/HICSS.2018.421.

75 G. A. Coppersmith, C. T. Harman, and M. H. Dredze, Measuring Post Traumatic Stress Disorder in Twitter, *Proc. 7th Int. AAAI Conf. Weblogs Soc. Media (ICWSM).*, vol. 2, no. 1, pp. 23–45, 2014.

76 B. Hao, L. Li, A. Li, and T. Zhu, Predicting Mental Health Status on Social Media, 2013, pp. 101–110. doi:10.1007/978-3-642-39137-8_12.

77 E. Saravia, C.-H. Chang, R. J. De Lorenzo, and Y.-S. Chen, MIDAS: Mental illness detection and analysis via social media, in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016, pp. 1418–1421. doi:10.1109/ASONAM.2016.7752434.

78 S. Chancellor, Z. Lin, E. Goodman, S. Zerwas, and M. De Choudhury, Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities, *Proc. 19th ACM Conf. Comput. Coop. Work Soc. Comput. (CSCW 2016)*, pp. 1171–1184, 2016. doi:10.1145/2818048.2819973.

79 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 - WSDM '17*, 2017, pp. 91–100. doi:10.1145/3018661.3018706.

80 S. Jamison-Powell, C. Linehan, L. Daley, A. Garbett, and S. Lawson, "I can't get no sleep": Discussing {#}insomnia on Twitter, *Proc. 2012 ACM Annu. Conf. Hum. Factors Comput. Syst. - CHI '12*, p. 1501, 2012. doi:10.1145/2207676.2208612.

81 S. R. Braithwaite, C. Giraud-Carrier, J. West, M. D. Barnes, and C. L. Hanson, Validating Machine Learning Algorithms for Twitter Data Against Established Measures of Suicidality, *JMIR Ment. Heal.*, vol. 3, no. 2, p. e21, May 2016. doi:10.2196/mental.4822. Medline:27185366.

82 G. Coppersmith, K. Ngo, R. Leary, and A. Wood, Exploratory Analysis of Social Media Prior to a Suicide Attempt, in *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, 2016, pp. 106–117. doi:10.18653/v1/W16-0311.

83 M. Lv, A. Li, T. Liu, and T. Zhu, Creating a Chinese suicide dictionary for identifying suicide risk on social media, *PeerJ*, vol. 3, p. e1455, 2015. doi:10.7717/peerj.1455. Medline:26713232.

84 B. O'Dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, and H. Christensen, Detecting suicidality on Twitter, *Internet Interv.*, vol. 2, no. 2, pp. 183–188, May 2015. doi:10.1016/j.invent.2015.03.005.

85 P. Burnap, W. Colombo, and J. Scourfield, Machine Classification and Analysis of Suicide-Related Communication on Twitter, *Proc. 26th ACM Conf.*

*Hypertext Soc. Media - HT '15*, pp. 75–84, 2015. doi:10.1145/2700171.2791023.

86 L. Zhang, X. Huang, T. Liu, A. Li, Z. Chen, and T. Zhu, Using Linguistic Features to Estimate Suicide Probability of Chinese Microblog Users, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8944, 2015, pp. 549–559. doi:10.1007/978-3-319-15554-8_45.

87 X. Huang, L. Zhang, D. Chiu, T. Liu, X. Li, and T. Zhu, Detecting Suicidal Ideation in Chinese Microblogs with Psychological Lexicons, *Proc. - 2014 IEEE Int. Conf. Ubiquitous Intell. Comput. 2014 IEEE Int. Conf. Auton. Trust. Comput. 2014 IEEE Int. Conf. Scalable Comput. Commun. Assoc. Sy*, pp. 844–849, 2015. doi:10.1109/UIC-ATC-ScalCom.2014.48.

88 C. Homan, R. Johar, T. Liu, M. Lytle, V. Silenzio, and C. Ovesdotter Alm, Toward Macro-Insights for Suicide Prevention: Analyzing Fine-Grained Distress at Scale, in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014, pp. 107–117. doi:10.3115/v1/W14-3213.

89 A. Roy, K. Nikolitch, R. McGinn, S. Jinah, W. Klement, and Z. A. Kaminsky, A machine learning approach predicts future risk to suicidal ideation from social media data, *npj Digit. Med.*, vol. 3, no. 1, p. 78, 2020. doi:10.1038/s41746-020-0287-6.

90 Y. Ma and Y. Cao, Dual Attention based Suicide Risk Detection on Social Media, in *2020 IEEE International Conference on Artificial Intelligence and*

*Computer Applications (ICAICA)*, 2020, pp. 637–640. doi:10.1109/ICAICA50127.2020.9182380.

91 A. Mbarek, S. Jamoussi, A. Charfi, and A. Ben Hamadou, Suicidal Profiles Detection in Twitter, in *Proceedings of the 15th International Conference on Web Information Systems and Technologies*, 2019, pp. 289–296. doi:10.5220/0008167602890296.

92 S. Fodeh, T. Li, K. Menczynski, T. Burgette, A. Harris, G. Ilita, S. Rao, J. Gemmell, and D. Raicu, Using Machine Learning Algorithms to Detect Suicide Risk Factors on Twitter, in *2019 International Conference on Data Mining Workshops (ICDMW)*, 2019, pp. 941–948. doi:10.1109/ICDMW.2019.00137.

93 G. Coppersmith, R. Leary, P. Crutchley, and A. Fine, Natural Language Processing of Social Media as Screening for Suicide Risk, *Biomed. Inform. Insights*, vol. 10, p. 117822261879286, Jan. 2018. doi:10.1177/1178222618792860.

94 D. Ramírez-Cifuentes, A. Freire, R. Baeza-Yates, J. Puntí, P. Medina-Bravo, D. A. Velazquez, J. M. Gonfaus, and J. Gonzàlez, Detection of Suicidal Ideation on Social Media: Multimodal, Relational, and Behavioral Analysis, *J. Med. Internet Res.*, vol. 22, no. 7, p. e17758, Jul. 2020. doi:10.2196/17758.

95 L. Guan, B. Hao, Q. Cheng, P. S. Yip, and T. Zhu, Identifying Chinese microblog users with high suicide probability using internet-based profile and linguistic features: classification model, *JMIR Ment. Heal.*, vol. 2, no. 2, pp. e17–e17, 2015. doi:10.2196/mental.4227. Medline:26543921.

96 H. A. Schwartz, M. Sap, M. L. Kern, J. C. Eichstaedt, A. Kapelner, M. Agrawal,

E. Blanco, L. Dziurzynski, G. Park, D. Stillwell, M. Kosinski, M. E. P. Seligman, and L. H. Ungar, Predicting individual well-being through the Language of social media, *Pac. Symp. Biocomput.*, vol. 21, pp. 516–527, 2016. Medline:26776214.

97 P. Liu, W. Tov, M. Kosinski, D. J. Stillwell, L. Qiu, Liu Pan, Tov William, Kosinski Michal, Stillwell David J, and Qiu Lin, Do Facebook Status Updates Reflect Subjective Well-Being?, *Cyberpsychology, Behav. Soc. Netw.*, vol. 18, no. 7, pp. 373–379, Jul. 2015. doi:10.1089/cyber.2015.0022.

98 A. O. Durahim and M. Coşkun, #iamhappybecause: Gross National Happiness through Twitter analysis and big data, *Technol. Forecast. Soc. Change*, vol. 99, pp. 92–105, 2015. doi:10.1016/j.techfore.2015.06.035.

99 C. Kuang, Z. Liu, M. Sun, F. Yu, and P. Ma, Quantifying Chinese happiness via large-scale microblogging data, *Proc. - 11th Web Inf. Syst. Appl. Conf. WISA 2014*, pp. 227–230, 2014. doi:10.1109/WISA.2014.48.

100 B. Hao, L. Li, R. Gao, A. Li, and T. Zhu, Sensing Subjective Well-being from Social Media, *Arxiv - Soc. Media Intell.*, vol. 8610, no. Chapter 27, pp. 324–335, Mar. 2014. doi:10.1007/978-3-319-09912-5_27.

101 H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, R. E. Lucas, M. Agrawal, G. J. Park, S. K. Lakshmikanth, S. Jha, M. E. P. Seligman, and L. H. Ungar, Characterizing geographic variation in well-being using tweets, in *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, 2013.

102 J. Bollen, B. Gonçalves, G. Ruan, and H. Mao, Happiness Is Assortative in

Online Social Networks, *Artif. Life*, vol. 17, no. 3, pp. 237–251, Jul. 2011. doi:10.1162/artl_a_00034. Medline:21554117.

103 D. Marengo, D. Azucar, C. Longobardi, and M. Settanni, Mining Facebook data for Quality of Life assessment, *Behav. Inf. Technol.*, pp. 1–11, Jan. 2020. doi:10.1080/0144929X.2019.1711454.

104 S. Volkova, K. Han, and C. Corley, Using Social Media to Measure Student Wellbeing: A Large-Scale Study of Emotional Response in Academic Discourse, 2016, pp. 510–526. doi:10.1007/978-3-319-47880-7_32.

105 V. L. Dos Reis and A. Culotta, Using Matched Samples to Estimate the Effects of Exercise on Mental Health from Twitter, in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, pp. 182–188.

106 A. Kittur, E. H. Chi, and B. Suh, Crowdsourcing user studies with Mechanical Turk, in *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*, 2008, p. 453. doi:10.1145/1357054.1357127.

107 M. Kosinski, S. C. Matz, S. D. Gosling, V. Popov, and D. Stillwell, Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines., *Am. Psychol.*, vol. 70, no. 6, pp. 543–56, Sep. 2015. doi:10.1037/a0039210. Medline:26348336.

108 S. Kiritchenko, X. Zhu, and S. M. Mohammad, Sentiment analysis of short informal texts, *J. Artif. Intell. Res.*, vol. 50, pp. 723–762, 2014. doi:10.1613/jair.4272.

109 T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C.

Cardie, E. Riloff, and S. Patwardhan, OpinionFinder: A System for Subjectivity Analysis, in *Proceedings of HLT/EMNLP on Interactive Demonstrations* -, 2005, no. October, pp. 34–35. doi:10.3115/1225733.1225751.

110 M. Thelwall, K. Buckley, G. Paltoglou, and D. Cai, Sentiment Strength Detection in Short Informal Text, *Am. Soc. Informational Sci. Technol.*, vol. 61, no. 12, pp. 2544–2558, 2010. doi:10.1002/asi. Medline:502955140.

111 D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent Dirichlet Allocation, *J. Mach. Learn. Res.*, vol. 1, no. 4–5, pp. 993–1022, 2003.

112 M. M. Bradley and P. J. Lang, Affective norms for English words (ANEW): Instruction manual and affective ratings, 1999.

113 M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. The MIT Press, 2012.

114 R. Sukthanker, S. Poria, E. Cambria, and R. Thirunavukarasu, Anaphora and coreference resolution: A review, *Inf. Fusion*, vol. 59, pp. 139–162, Jul. 2020. doi:10.1016/j.inffus.2020.01.010.

115 G. A. Miller, WordNet: a lexical database for English, *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995. doi:10.1145/219717.219748.

116 A. K. McCallum, MALLET: A Machine Learning for Language Toolkit, 2002.

117 K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, Feature-rich part-of-speech tagging with a cyclic dependency network, in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL ’03*, 2003, vol. 1, pp. 173–180. doi:10.3115/1073445.1073478.

118 D. Castelvecchi, Can we open the black box of AI?, *Nature*, vol. 538, no. 7623, pp. 20–23, Oct. 2016. doi:10.1038/538020a.

119 V. Kotu and B. Deshpande, Data Science Process, in *Data Science*, Elsevier, 2019, pp. 19–37. doi:10.1016/B978-0-12-814761-0.00002-2.

120 J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in *Proceedings of the 2019 Conference of the North*, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

121 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.

122 D. Kotzias, M. Denil, N. de Freitas, and P. Smyth, From Group to Individual Labels Using Deep Features, in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, 2015, pp. 597–606. doi:10.1145/2783258.2783380.

123 C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

124 N. Jakob and I. Gurevych, Using Anaphora Resolution to Improve Opinion Target Identification in Movie Reviews, in *Proceedings of the {ACL} 2010 Conference Short Papers*, 2010, pp. 263–268.

125 K. Lee, L. He, M. Lewis, and L. Zettlemoyer, End-to-end Neural Coreference Resolution, in *Proceedings of the 2017 Conference on Empirical Methods in*

*Natural Language Processing*, 2017, pp. 188–197. doi:10.18653/v1/D17-1018.

126 K. Lee, L. He, and L. Zettlemoyer, Higher-Order Coreference Resolution with Coarse-to-Fine Inference, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, pp. 687–692. doi:10.18653/v1/N18-2108.

127 R. Othman, R. Belkaroui, and R. Faiz, Extracting Product Features for Opinion Mining Using Public Conversations in Twitter, *Procedia Comput. Sci.*, vol. 112, pp. 927–935, 2017. doi:10.1016/j.procs.2017.08.122.

128 S. A. Moorhead, D. E. Hazlett, L. Harrison, J. K. Carroll, A. Irwin, and C. Hoving, A New Dimension of Health Care: Systematic Review of the Uses, Benefits, and Limitations of Social Media for Health Communication, *J. Med. Internet Res.*, vol. 15, no. 4, p. e85, Apr. 2013. doi:10.2196/jmir.1933.

129 D. Scanfeld, V. Scanfeld, and E. L. Larson, Dissemination of health information through social networks: Twitter and antibiotics, *Am. J. Infect. Control*, vol. 38, no. 3, pp. 182–188, Apr. 2010. doi:10.1016/j.ajic.2009.11.004.

130 D. P. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, *CoRR*, vol. abs/1412.6, 2014.

131 D. J. Stillwell and M. Kosinski, myPersonality project: Example of successful utilization of online social networks for large-scale social research, in *Proceedings of the 1st ACM Workshop on Mobile Systems for Computational Social Science (MobiSys)*, 2012.

132 J. Wang, Z. Wang, D. Zhang, and J. Yan, Combining Knowledge with Deep

Convolutional Neural Networks for Short Text Classification, in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 2915–2921.

133 A. K. Uysal and Y. L. Murphey, Sentiment Classification: Feature Selection Based Approaches Versus Deep Learning, in *2017 IEEE International Conference on Computer and Information Technology (CIT)*, 2017, pp. 23–30. doi:10.1109/CIT.2017.53.

134 J. Y. Lee and F. Dernoncourt, Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks, in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 515–520. doi:10.18653/v1/N16-1062.

135 S. Lai, L. Xu, K. Liu, and J. Zhao, Recurrent convolutional neural networks for text classification, in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

136 X. Zhang, J. Zhao, and Y. LeCun, Character-level Convolutional Networks for Text Classification, in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 649–657.

137 P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification, *Neurocomputing*, vol. 174, pp. 806–814, Jan. 2016. doi:10.1016/j.neucom.2015.09.096.

138 W. Yin, K. Kann, M. Yu, and H. Schütze, Comparative Study of CNN and RNN for Natural Language Processing, Feb. 2017.

139 Y. Kim, Convolutional Neural Networks for Sentence Classification, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746–1751. doi:10.3115/v1/D14-1181.

140 M. Makita, A. Mas-Bleda, S. Morris, and M. Thelwall, Mental Health Discourses on Twitter during Mental Health Awareness Week, *Issues Ment. Health Nurs.*, vol. 42, no. 5, pp. 437–450, May 2021. doi:10.1080/01612840.2020.1814914.

141 N. Berry, F. Lobban, M. Belousov, R. Emsley, G. Nenadic, and S. Bucci, #WhyWeTweetMH: Understanding Why People Use Twitter to Discuss Mental Health Problems, *J. Med. Internet Res.*, vol. 19, no. 4, p. e107, Apr. 2017. doi:10.2196/jmir.6173.

142 X. Yang, R. McCreadie, C. Macdonald, and I. Ounis, Transfer Learning for Multi-language Twitter Election Classification, in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 2017, pp. 341–348. doi:10.1145/3110025.3110059.

143 N. Gopalkrishnan, Cultural Diversity and Mental Health: Considerations for Policy and Practice, *Front. Public Heal.*, vol. 6, Jun. 2018. doi:10.3389/fpubh.2018.00179.

144 Y. Jang, C.-H. Park, and Y.-S. Seo, Fake News Analysis Modeling Using Quote Retweet, *Electronics*, vol. 8, no. 12, p. 1377, Nov. 2019.

doi:10.3390/electronics8121377.

145 N. Majumder, R. Bhardwaj, S. Poria, A. Gelbukh, and A. Hussain, Improving aspect-level sentiment analysis with aspect extraction, *Neural Comput. Appl.*, Aug. 2020. doi:10.1007/s00521-020-05287-7.

146 A. D. I. Kramer, J. E. Guillory, and J. T. Hancock, Experimental evidence of massive-scale emotional contagion through social networks, *Proc. Natl. Acad. Sci.*, vol. 111, no. 24, pp. 8788–8790, Jun. 2014. doi:10.1073/pnas.1320040111.

147 I. M. Verma, Editorial expression of concern: Experimental evidence of massive-scale emotional contagion through social networks., *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 29, p. 10779, Jul. 2014. doi:10.1073/pnas.1412469111. Medline:24994898.

148 K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis, Tastes, ties, and time: A new social network dataset using Facebook.com, *Soc. Networks*, vol. 30, no. 4, pp. 330–342, Oct. 2008. doi:10.1016/j.socnet.2008.07.002.

149 M. Zimmer, “But the data is already public”: on the ethics of research in Facebook, *Ethics Inf. Technol.*, vol. 12, no. 4, pp. 313–325, Dec. 2010. doi:10.1007/s10676-010-9227-5.

150 M. Sykora, S. Elayan, N. Barbour, and T. Jackson, A Survey of the Ethics of Social Media Analytics, in *ECSM-2020 - 7th European Conference on Social Media*, 2020. doi:10.34190/ESM.20.047.

151 E. Ford, K. Curlewis, A. Wongkoblap, and V. Curcin, Caring or creepy? A

mixed-methods survey of public opinions on using social media content for identifying users with depression and targeting mental health-care advertising, *JMIR Ment. Heal.*, Aug. 2019.

152 J. Taylor and C. Pagliari, Mining social media data: How are research sponsors and researchers addressing the ethical challenges?, *Res. Ethics*, vol. 14, no. 2, pp. 1–39, Apr. 2018. doi:10.1177/1747016117738559.

153 S. Golder, S. Ahmed, G. Norman, and A. Booth, Attitudes toward the ethics of research using social media: A systematic review, *J. Med. Internet Res.*, vol. 19, no. 6, 2017. doi:10.2196/jmir.7082. Medline:28588006.

154 A. Benton, G. Coppersmith, and M. Dredze, Ethical Research Protocols for Social Media Health Research, in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 2017, pp. 94–102. doi:10.18653/v1/W17-1612.

155 M. Fuller, Big data and the Facebook scandal: Issues and responses, *Theology*, vol. 122, no. 1, pp. 14–21, 2019. doi:10.1177/0040571X18805908.

156 A. Benton, G. Coppersmith, and M. Dredze, Ethical Research Protocols for Social Media Health Research, in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 2017, pp. 94–102. doi:10.18653/v1/W17-1612.

157 A. Richterich, *The Big Data Agenda: Data Ethics and Critical Data Studies*. University of Westminster Press, 2018. doi:10.16997/book14.

158 W. Youyou, M. Kosinski, and D. Stillwell, Computer-based personality judgments are more accurate than those made by humans, *Proc. Natl. Acad.*

*Sci.*, vol. 112, no. 4, pp. 1036–1040, Jan. 2015. doi:10.1073/pnas.1418680112.

159 J. Mikal, S. Hurst, and M. Conway, Ethical issues in using Twitter for population-level depression monitoring: a qualitative study, *BMC Med. Ethics*, vol. 17, no. 1, p. 22, Dec. 2016. doi:10.1186/s12910-016-0105-5.

160 K. Shaughnessy, R. Reyes, K. Shankardass, M. Sykora, R. Feick, H. Lawrence, and C. Robertson, Using geolocated social media for ecological momentary assessments of emotion: Innovative opportunities in psychology science and practice., *Can. Psychol. Can.*, vol. 59, no. 1, pp. 47–53, Feb. 2018. doi:10.1037/cap0000099.

161 J. Taylor and C. Pagliari, Mining social media data: How are research sponsors and researchers addressing the ethical challenges?, *Res. Ethics*, vol. 14, no. 2, pp. 1–39, Apr. 2018. doi:10.1177/1747016117738559.

162 A. Richterich, *The Big Data Agenda: Data Ethics and Critical Data Studies*. University of Westminster Press, 2018. doi:10.16997/book14.

163 M. Conway, Ethical issues in using Twitter for public health surveillance and research: developing a taxonomy of ethical concepts from the research literature., *J. Med. Internet Res.*, vol. 16, no. 12, p. e290, Dec. 2014. doi:10.2196/jmir.3617. Medline:25533619.

164 R. McKee, Ethical issues in using social media for health and health care research, *Health Policy (New. York).*, vol. 110, no. 2–3, pp. 298–301, May 2013. doi:10.1016/j.healthpol.2013.02.006.

165 D. Wilkinson and M. Thelwall, Researching Personal Information on the Public Web: Methods and Ethics, *Soc. Sci. Comput. Rev.*, vol. 29, no. 4, pp. 387–401,

Nov. 2011. doi:10.1177/0894439310378979.

166 C. A. Sula, Research Ethics in an Age of Big Data, *Bull. Assoc. Inf. Sci. Technol.*, vol. 42, no. 2, pp. 17–21, Jan. 2016. doi:10.1002/bul2.2016.1720420207.

167 A. Markham and E. Buchanan, Ethical Decision-Making and Internet Research:Version 2.0, *Association of Internet Researchers*, 2012. [Online]. Available: http://aoir.org/reports/ethics2.pdf: . Webcite: 6hhZbLexD.

168 J. Oates, R. Kwiatkowski, and L. Morrison-Coulthard, Code of human research ethics, *Br. Psychol. Soc.*, 2010.

169 C. Fiesler, N. Beard, and B. C. Keegan, No Robots, Spiders, or Scrapers: Legal and Ethical Regulation of Data Collection Methods in Social Media Terms of Service, *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 14, no. 1, pp. 187–196, May 2020.

170 E. O’Neil, API Updates and Important Changes, *Facebook*, 2019. [Online]. Available: https://developers.facebook.com/blog/post/2019/04/25/api-updates/. [Accessed: 17-Mar-2020].

171 D. Oh, S. Elayan, M. Sykora, and J. Downey, Unpacking uncivil society: Incivility and intolerance in the 2018 Irish abortion referendum discussions on Twitter, *Nord. Rev.*, vol. 42, no. s1, pp. 103–118, Mar. 2021. doi:10.2478/nor-2021-0009.

172 M. Hughes, I. Li, S. Kotoulas, and T. Suzumura, Medical Text Classification Using Convolutional Neural Networks., *Stud. Health Technol. Inform.*, vol. 235, pp. 246–250, 2017. Medline:28423791.

173 S. S. Yadav and S. M. Jadhav, Deep convolutional neural network based medical image classification for disease diagnosis, *J. Big Data*, vol. 6, no. 1, p. 113, 2019. doi:10.1186/s40537-019-0276-2.

174 G. Montavon, W. Samek, and K.-R. Müller, Methods for interpreting and understanding deep neural networks, *Digit. Signal Process.*, vol. 73, pp. 1–15, Feb. 2018. doi:10.1016/j.dsp.2017.10.011.

175 R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, Explainable Machine Learning for Scientific Insights and Discoveries, *IEEE Access*, vol. 8, pp. 42200–42216, 2020. doi:10.1109/ACCESS.2020.2976199.

176 M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You?” Explaining the Predictions of Any Classifier, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144. doi:10.1145/2939672.2939778.

177 S. M. Lundberg and S.-I. Lee, A Unified Approach to Interpreting Model Predictions, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4768–4777.

178 P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, Explainable AI: A Review of Machine Learning Interpretability Methods, *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020. doi:10.3390/e23010018.

179 F. A. Huppert, Psychological Well-being: Evidence Regarding its Causes and Consequences, *Appl. Psychol. Heal. Well-Being*, vol. 1, no. 2, pp. 137–164, Jul. 2009. doi:10.1111/j.1758-0854.2009.01008.x.

180 O. F.Y, A. J.E.T, A. O, H. J. O, O. O, and A. J, Supervised Machine Learning

Algorithms: Classification and Comparison, *Int. J. Comput. Trends Technol.*, vol. 48, no. 3, pp. 128–138, Jun. 2017. doi:10.14445/22312803/IJCTT-V48P126.

181 R. Arora and S. Suman, Comparative Analysis of Classification Algorithms on Different Datasets using WEKA, *Int. J. Comput. Appl.*, vol. 54, no. 13, pp. 21–25, Sep. 2012. doi:10.5120/8626-2492.

182 A. Singh, M. N., and R. Lakshmiganthan, Impact of Different Data Types on Classifier Performance of Random Forest, Naïve Bayes, and K-Nearest Neighbors Algorithms, *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 12, 2017. doi:10.14569/IJACSA.2017.081201.

183 Y. Pershad, P. Hangge, H. Albadawi, and R. Oklu, Social Medicine: Twitter in Healthcare, *J. Clin. Med.*, vol. 7, no. 6, p. 121, May 2018. doi:10.3390/jcm7060121.

184 U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, and P. Eckersley, Explainable Machine Learning in Deployment, Sep. 2019.

185 A. Holzinger, P. Kieseberg, E. Weippl, and A. M. Tjoa, Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI, 2018, pp. 1–8. doi:10.1007/978-3-319-99740-7_1.

186 A. Farseev and T.-S. Chua, TweetFit: Fusing Multiple Social Media and Sensor Data for Wellness Profile Learning, in *AAAI*, 2017.

187 A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, What do we need to

build explainable AI systems for the medical domain?, Dec. 2017.

188 L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, Explaining Explanations: An Overview of Interpretability of Machine Learning, May 2018.

189 M. P. Tully, L. Hassan, M. Oswald, and J. Ainsworth, Commercial use of health data-A public “trial” by citizens’ jury, *Learn. Heal. Syst.*, vol. 3, no. 4, p. e10200, Oct. 2019. doi:10.1002/lrh2.10200.

190 World Health Organization, Mental health: strengthening our response, 2018. [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/mental-health-strengthening-our-response.

191 World Health Organization, Mental health: strengthening mental health promotion, 2007. [Online]. Available: https://mindyourmindproject.org/wp-content/uploads/2014/11/WHO-Statement-on-Mental-Health-Promotion.pdf.

192 World Health Organization, A public health approach to mental health, in *The world health report 2001 - Mental Health: New Understanding, New Hope*, 2001.

193 National Collaborating Centre for Mental Health (UK), *Common Mental Health Disorders: Identification and Pathways to Care*. Leicester (UK): British Psychological Society, 2011.

194 S. McManus, P. Bebbington, R. Jenkins, T. Brugha, and (eds.), Mental health and wellbeing in England: Adult Psychiatric Morbidity Survey 2014, *Leeds: NHS Digital*, 2016. [Online]. Available: http://content.digital.nhs.uk/catalogue/PUB21748/apms-2014-full-rpt.pdf.

195 G. Vilagut, C. G. Forero, G. Barbaglia, and J. Alonso, Screening for Depression in the General Population with the Center for Epidemiologic Studies Depression (CES-D): A Systematic Review with Meta-Analysis, *PLoS One*, vol. 11, no. 5, p. e0155431, May 2016. doi:10.1371/journal.pone.0155431.

196 C. R. Cloninger, D. M. Svrakic, and T. R. Przybeck, Can personality assessment predict future depression? A twelve-month follow-up of 631 subjects, *J. Affect. Disord.*, vol. 92, no. 1, pp. 35–44, May 2006. doi:10.1016/j.jad.2005.12.034.

197 S. S. Rude, C. R. Valdez, S. Odom, and A. Ebrahimi, Negative Cognitive Biases Predict Subsequent Depression, *Cognit. Ther. Res.*, vol. 27, no. 4, pp. 415–429, 2003. doi:10.1023/A:1025472413805.

198 M. S. Robinson and L. B. Alloy, Negative Cognitive Styles and Stress-Reactive Rumination Interact to Predict Depression: A Prospective Study, *Cognit. Ther. Res.*, vol. 27, no. 3, pp. 275–291, 2003. doi:10.1023/A:1023914416469.

199 S. Rude, E.-M. Gortner, and J. Pennebaker, Language use of depressed and depression-vulnerable college students, *Cogn. Emot.*, vol. 18, no. 8, pp. 1121–1133, Dec. 2004. doi:10.1080/02699930441000030.

200 Institute for Health Metrics and Evaluation, Global Burden of Disease 2019, 2019. [Online]. Available: http://ghdx.healthdata.org/gbd-results-tool.

201 D. Razzouk, Burden and Indirect Costs of Mental Disorders, in *Mental Health Economics*, Cham: Springer International Publishing, 2017, pp. 381–391. doi:10.1007/978-3-319-55266-8_25.

202 S. Trautmann, J. Rehm, and H. Wittchen, The economic costs of mental disorders, *EMBO Rep.*, vol. 17, no. 9, pp. 1245–1249, Sep. 2016.

doi:10.15252/embr.201642951.

203 H. Ritchie and M. Roser, Mental Health, *Our World Data*, 2018.

204 J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre, Social media? Get serious! Understanding the functional building blocks of social media, *Bus. Horiz.*, vol. 54, no. 3, pp. 241–251, May 2011. doi:10.1016/j.bushor.2011.01.005.

205 T. Aichner and F. Jacob, Measuring the Degree of Corporate Social Media Use, *Int. J. Mark. Res.*, vol. 57, no. 2, pp. 257–276, Mar. 2015. doi:10.2501/IJMR-2015-018.

206 Y.-Q. Zhu and H.-G. Chen, Social media and human need satisfaction: Implications for social media marketing, *Bus. Horiz.*, vol. 58, no. 3, pp. 335–345, May 2015. doi:10.1016/j.bushor.2015.01.006.

207 Statista, Number of social media users worldwide from 2010 to 2020 (in billions), 2016. [Online]. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/. [Accessed: 25-Jan-2017].

208 D. P. Schultz, The human subject in psychological research., *Psychological Bulletin*, vol. 72, no. 3. American Psychological Association, US, pp. 214–228, 1969. doi:10.1037/h0027880.

209 World Medical Association, World Medical Association Declaration of Helsinki, *JAMA*, vol. 310, no. 20, p. 2191, Nov. 2013. doi:10.1001/jama.2013.281053.

210 A. Z. Klein, A. Sarker, H. Cai, D. Weissenbacher, and G. Gonzalez-Hernandez,

Social media mining for birth defects research: A rule-based, bootstrapping approach to collecting data for rare health-related events on Twitter, *J. Biomed. Inform.*, vol. 87, pp. 68–78, Nov. 2018. doi:10.1016/j.jbi.2018.10.001.

211 L. Deng and Y. Liu, Eds., *Deep Learning in Natural Language Processing*. Singapore: Springer Singapore, 2018. doi:10.1007/978-981-10-5209-5.

212 H. Dalianis, *Clinical Text Mining*. Cham: Springer International Publishing, 2018. doi:10.1007/978-3-319-78503-5.

213 Y. R. Tausczik and J. W. Pennebaker, The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods, *J. Lang. Soc. Psychol.*, vol. 29, no. 1, pp. 24–54, Mar. 2010. doi:10.1177/0261927X09351676.

214 J. Pennington, R. Socher, and C. D. Manning, GloVe: Global Vectors for Word Representation, in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

215 R. Mitkov, *Anaphora Resolution*, vol. 1. Oxford University Press, 2012. doi:10.1093/oxfordhb/9780199276349.013.0014.

216 I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

217 T. M. Mitchell, *Machine Learning*, 1st ed. New York, NY, USA: McGraw-Hill, Inc., 1997.

218 V. Chaoji, A. Hoonlor, and B. K. Szymanski, Recursive data mining for role identification, in *Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology - CSTST '08*, 2008, p. 218. doi:10.1145/1456223.1456270.

219 X. Zhu and A. B. Goldberg, Introduction to Semi-Supervised Learning, *Synth. Lect. Artif. Intell. Mach. Learn.*, vol. 3, no. 1, pp. 1–130, Jan. 2009. doi:10.2200/S00196ED1V01Y200906AIM006.

220 J. D. Keeler, D. E. Rumelhart, and W. K. Leow, Integrated Segmentation and Recognition of Hand-Printed Numerals, in *Advances in Neural Information Processing Systems 3*, R. P. Lippmann, J. E. Moody, and D. S. Touretzky, Eds. Morgan-Kaufmann, 1991, pp. 557–563.

221 D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning representations by back-propagating errors, *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986. doi:10.1038/323533a0.

222 Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. doi:10.1109/5.726791.

223 Y. Le Cun, L. D. Jackel, B. Boser, J. S. Denker, H. P. Graf, I. Guyon, D. Henderson, R. E. Howard, and W. Hubbard, Handwritten digit recognition: applications of neural network chips and automatic learning, *IEEE Commun. Mag.*, vol. 27, no. 11, pp. 41–46, Nov. 1989. doi:10.1109/35.41400.

224 K. Fukushima, Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, Apr. 1980. doi:10.1007/BF00344251.

225 S. Hochreiter and J. Schmidhuber, Long Short-Term Memory, *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. doi:10.1162/neco.1997.9.8.1735.

226 F. Chollet, *Deep Learning with Python*. Manning Publications Company, 2017.

227 T. Fawcett, An introduction to ROC analysis, *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006. doi:10.1016/j.patrec.2005.10.010.

228 J. A. Hanley and B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve., *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982. doi:10.1148/radiology.143.1.7063747. Medline:7063747.

229 S. Hauenstein, S. N. Wood, and C. F. Dormann, Computing AIC for black-box models using generalized degrees of freedom: A comparison with cross-validation, *Commun. Stat. - Simul. Comput.*, vol. 47, no. 5, pp. 1382–1396, May 2018. doi:10.1080/03610918.2017.1315728.

230 S. I. Vrieze, Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC)., *Psychol. Methods*, vol. 17, no. 2, pp. 228–243, 2012. doi:10.1037/a0027127.

231 H. Akaike, A new look at the statistical model identification, *IEEE Trans. Automat. Contr.*, vol. 19, no. 6, pp. 716–723, Dec. 1974. doi:10.1109/TAC.1974.1100705.

232 G. Panchal, A. Ganatra, Y. P. Kosta, and D. Panchal, Searching Most Efficient Neural Network Architecture Using Akaike’s Information Criterion (AIC), *Int. J. Comput. Appl.*, vol. 975, p. 8887, 2010.

233 C. M. HURVICH and C.-L. TSAI, Regression and time series model selection in small samples, *Biometrika*, vol. 76, no. 2, pp. 297–307, 1989. doi:10.1093/biomet/76.2.297.

# Appendix A
# Background

This chapter provides the necessary background to the novel research work presented in later chapters. We start off by detailing types of mental disorders, mental illness screening, burden of mental diseases, and depression. This is followed by approaches to social media data collection, data annotation and the building blocks for predictive modelling. Finally, different techniques for model evaluation are described.

## A.1 Mental Health Disorders

World Health Organization (WHO) defines "*mental health* as a state of well-being in which the individual realises his or her abilities, can cope with the normal stresses of life, work productively and fruitfully, and is able to make a contribution to his or her community" [190, 191]. While an individual might periodically feel deeply unsatisfied, people suffering from being constantly and deeply unhappy may fall into mental illness. Mental health disorder is characterised as a medical condition in which a patient tends to express negative emotions, thinking, behaviours, and relationships [192].

### A.1.1 Types of Mental Illness

Mental illness comprises different types of disorders. Common mental disorders include depression, generalised anxiety disorder (GAD), social anxiety disorder, panic disorder, phobias, obsessive-compulsive disorder (OCD), and post-traumatic stress disorder (PTSD) [193, 194].

*Depression*: It refers to a situation in which people lose interest, remain in unstable moods, and socially isolate from others. Sometimes they might experience physical problems in parallel (e.g., weight loss/gain and poor sleep).

*Generalised anxiety disorder (GAD)*: Patients with GAD suffer extreme and unnecessary levels of anxiety and worry about activities and events of their daily life. They cannot easily control the anxiety and worry, occurring over at least 6 months.

*Social anxiety disorder*: This is confined to one specific phobia, namely social phobia. An individual with this disorder tends to avoid social situations, because of a fear of being distinguished or denied, usually leading to problems in education, employment, and daily life's activities.

*Panic disorder*: Panic disorder refers to a condition in which a person fears unexpected and immediate particular situations, resulting in the person avoiding them.

*Phobias*: A person will feel extremely afraid of specific objects or situations, leading to avoid them.

*Obsessive-compulsive disorder (OCD)*: Common signs of OCD are obsessions and compulsions. People suffering from this disorder experience overthinking and/or perform a repetitive pattern uncontrollably.

*Post-traumatic stress disorder (PTSD)*: PTSD is characterised by recurrent thoughts and fear of traumatic events (e.g., violence, accidents, disasters, or wars).

## A.1.2 Mental Health Screening

To screen symptoms of these disorders and diagnose them, since April 2006, General Practitioners (GP) have used well-established and widely validated questionnaires. Additionally, GPs might consider a patient’s medical history rather than relying exclusively on score of the questionnaires [193].

*The Centre for Epidemiological Study Depression Scale (CES-D)* is a common and well-designed questionnaire to measure and screen depressive tendencies [9]. It has been used to screen respondents with depression both in the general population and in primary care settings [195]. This questionnaire consists of 20 items. Each of them asks how often the respondent has felt in different states or moods over the last week. The answers are rarely or none of the time (less than 1 day), some or a little of the time (1-2 days), occasionally or a moderate amount of the time (3-4 days), and most or all of the time (5-7 days). The total score ranges from 0 to 60 and can used to screen a respondent as depressed or non-depressed.

Apart from the usage of surveys to diagnose mental illness, personality assessment [196], negative cognitive biases [197], ruminative thinking [198], and linguistic analysis [199] can also predict the prevalence of mental disorders [24].

## A.1.3 Burden of Mental Disorders

Mental illness is one of the leading diseases for global population. Figure A-1 shows disability-adjusted life years (DALYs – is measured from years of life lost and years lived with disability – as displayed in Figure A-2) caused by diseases between 1990 and 2019. As can be seen, mental and substance abuse disorders are constantly growing from ranked ninth in 1990 and to fourth ranked cause of DALYs in 2019 globally. It tends to continuously increase afterward. The number of global patients suffering from mental disorders is estimated to be over 970 million in 2019 [200].

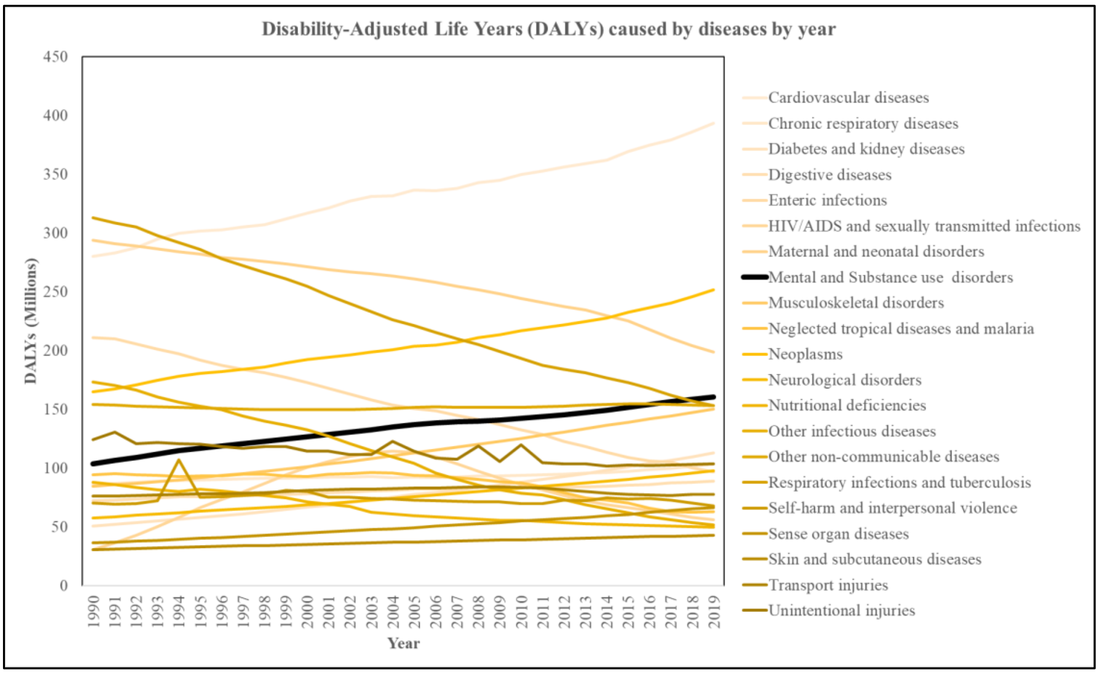


*Figure A-1 Disability-Adjusted Life Years (DALYs) caused by global diseases from 1990 to 2019 (data from* [200]*)*

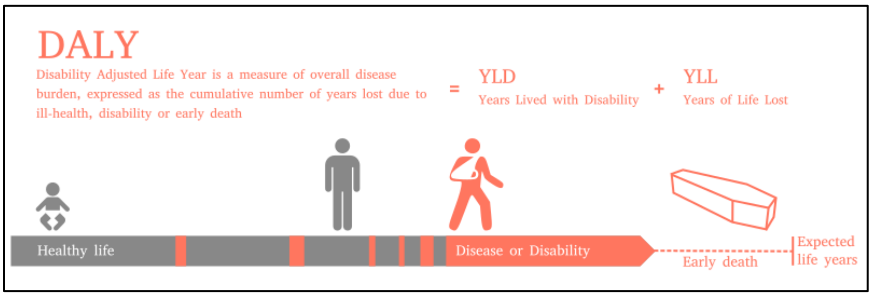

DALY
Disability Adjusted Life Year is a measure of overall disease burden, expressed as the cumulative number of years lost due to ill-health, disability or early death
=
YLD
Years Lived with Disability
+
YLL
Years of Life Lost
Healthy life
Disease or Disability
Early death
Expected life years

*Figure A-2 DALY disability affected life year infographic retrieved from Wikimedia Commons*[5]

[5] https://commons.wikimedia.org/wiki/File:DALY_disability_affected_life_year_info-graphic.svg

In terms of an economic burden, the global costs of mental health problems approximated 2.5 trillion USD in 2010, and it is expected to be 6.1 trillion USD by 2030 (see Figure A-3-A). It disrupts economic growth in many countries, especially in low- and middle-income. The calculation of the costs normally takes no account of indirect costs such as *national level*: mental, human capital and economic growth losses; *workplace level*: work productivity losses, worker replacement costs, and earlier retirement; and *individual level*: income losses and poverty, poor educational attainment, and lower life expectancy [3, 201]. Figure A-3-A indicates that the indirect costs of mental illness are greater than the direct costs. Mental disorder costs also contribute to the largest proportion of economic losses among the leading diseases [202] as shown in Figure A-3-B and C.

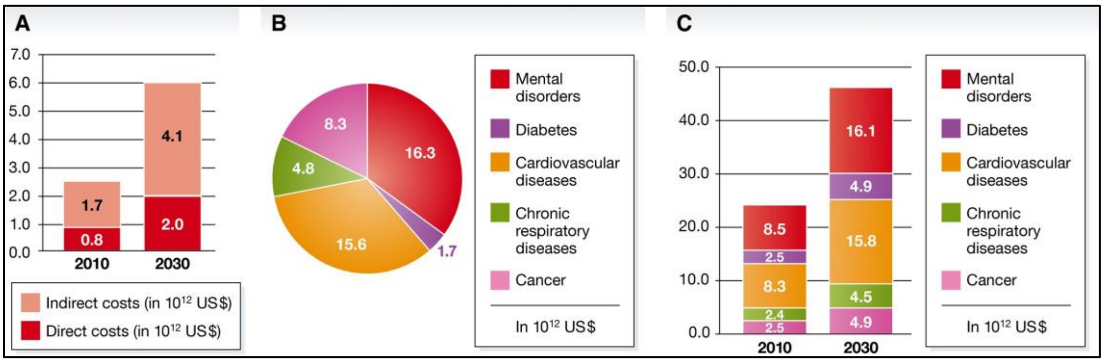


*Figure A-3 Economic costs of mental disorders in trillion US$ using three different approaches: direct and indirect costs of mental disorders (A), the cumulative economic output loss between 2011 and 2030 (B), and risks and money to quantify the risk of disability or death (C) (taken from* [202]*)*

### A.1.4 Depression

Depression contributes to the most DALYs (see Figure A-4). According to the statistics of WHO report, the number of global patients suffering from depression is estimated to be over 264 million in 2020 [1]. Even with so many people suffering from depression, 93% of them received no treatment in developing countries and 72% could not access mental health services in developed countries. Scaled-up treatment for depression can reduce the loss of 230 billion USD [7].

Depression is a common mental disorder ranging from mild to severe. In the worst case, patients with severe depressive disorder can commit suicide. People with depression tended to commit suicide 20 times more often than people without depression [203].

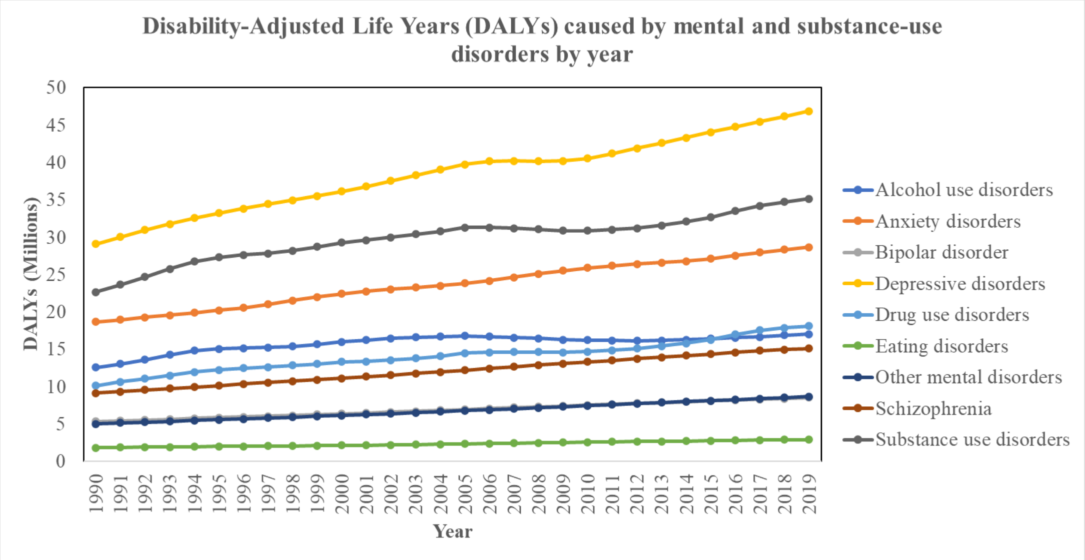


*Figure A-4 Disability-Adjusted Life Years (DALYs) caused by mental and substance-use diseases from 1990 to 2019 (data from* [200]*)*

The above information shows that mental disorders, especially depression, are among the top-5 most concerning world health issues to be addressed. This highlights the need of a novel and effective approach for early detecting people with depression and broader providing better mental health services accessible by everyone.

## A.2 Social Media Data

***Social media*** are online platforms that allow users to create, discuss, modify, and exchange content. Users registering on such platform can present their identities to others, communicate with others, form a wide variety of interest groups, and establish or maintain relationships [204].

Social media comprises various types of platforms with different characteristics. This helps define a profile-based platform to collect users' data and annotate those users associated with mental disorders.

### A.2.1 Social Media Typology

Based on the above definition, social media can be classified into the 13 different types. Table A-1 describes the definitions of each social media [205].

*Table A-1 Definitions of social media types*

| Type of Social Media | Description |
|---|---|
| *Blogs* | A blog, shortened from web and log, is a list of ordered posts published by a poster. Other users can read and comment. |

| **Type of Social Media** | **Description** |
|---|---|
| | *Example: wordpress.com, medium.net* |
| *Business networks* | A business network is a platform where individuals create a professional profile to connect and maintain professional contacts. Companies seek new employees based on professional profiles.<br>*Example: linkedin.com, xing.com* |
| *Collaborative projects* | A platform facilitates users with a common interest to collaborate to develop a new project. The result from the project is typically released as open source.<br>*Example: wikipedia.org, mozilla.org* |
| *Enterprise social networks* | Enterprise social network is a private platform for employees in a specific company to connect to one another and exchange knowledge in the company.<br>*Example: yammer.com, socialcast.com* |
| *Forums* | A forum is an online discussion board where users can create topics to ask or exchange with other users and reply to user's topics.<br>*Example: quora.com, reddit.com* |

| **Type of Social Media** | **Description** |
|---|---|
| *Microblogs* | Microblogs are like blogs but there is a word limitation of the length of posting e.g., 280 characters. Posts may contain pictures, videos, and URLs. Users can follow other accounts.<br>*Example: twitter.com, weibo.com* |
| *Photo sharing* | Photo sharing is a platform where users can upload, caption and share photos. Other users can comment the shared photos.<br>*Example: flickr.com, instagram.com* |
| *Products/ services review* | Product and service review platforms are websites where users can evaluate products and write or read reviews. The sites often sell or provide product information.<br>*Example: amazon.com, yelp.com* |
| *Social bookmarking* | Social bookmarking is a centralised platform for users to create and arrange bookmarks in order to share with others.<br>*Example: delicious.com, pinterest.com* |
| *Social gaming* | Social gaming refers to online games that users can interact and collaborate with other players.<br>*Example: gameloft.com, steam.com* |

| **Type of Social Media** | **Description** |
|---|---|
| *Social networks* | Social networking platforms allow users to create profiles and connect to others who know each other or have common interests. Users can post text, pictures, videos, and URLs.<br>*Example: facebook.com, vk.com* |
| *Video sharing* | Video sharing refers to an online platform where users can upload and share videos. The platform may allow users to comment on the videos.<br>*Example: youtube.com, tiktok.com* |
| *Virtual worlds* | Virtual worlds are online places where users can create a customised avatar to represent themselves. These avatars are used to interact and communicate with other avatars.<br>*Example: secondlife.com, twinity.com* |

Another social media typology is based on the nature of connection (profile-based versus content-based) and the level of customisation of messages (broadcast versus customised). Table A-2 provides details of nature of connection. Profile-based versus content-based and customised versus broadcast dimensions can be used to define four types of social media using a two-by-two matrix, as shown in Table A-3 [206].

*Table A-2 Profile-based vs. content-based social media (taken from* [206]*)*

| | **Profile-based** | **Content-based** |
|---|---|---|
| **Focal point** | The individual member | Contents posted |
| **Nature of information** | Topics are typically related to the person | Discussions and comments are based around contents posted |
| **Main purpose** | Users make connections mainly because they are interested in the user behind the profile | Users make connections because they like the contents a certain profile provides |
| **Examples** | Facebook, Twitter, Line, Whatsapp | Flickr, Instagram, Pinterest, YouTube |

*Table A-3 Social media matrix (taken from* [206]*)*

| | *Customised Message* | *Broadcast Message* |
|---|---|---|
| *Profile-based* | **Relationship**<br>Allowing users to connect, reconnect, communicate, and build relationships.<br>(e.g., Facebook, LinkedIn, Line, Whatsapp) | **Self-Media**<br>Allowing users to broadcast their updates and others to follow.<br>(e.g., Twitter, Weibo) |
| *Content-based* | **Collaboration**<br>Allowing users to collaboratively find answers, advice, help, and reach consensus.<br>(e.g., Quora, Reddit, Yahoo! Answers) | **Creative outlets**<br>Allowing users to share their interest, creativity, and hobbies with each other.<br>(e.g., YouTube, Flickr, Foodily, Pinterest) |

The focus of this thesis is on individuals' content-based and profile-based social media platforms, because we explore mental illness content on users' profiles and detect users with mental disorder from their textual content. Profile-based social media platforms selected for this study are microblogs like *Twitter* and social networks like *Facebook*. Unlike other types of social media, social networking and microblogging platforms allow users to create personal profiles, establish new relationships as well as maintain close friendships, publish status updates, and have interactions with other users. Those users openly express a variety of thoughts, feelings, and emotions every day.

In 2020 there are over 3.6 billion social media users [207]. Just the most popular social networking site, Facebook, contains over 1.88 billion daily active users in March 2021 [13]. The most popular microblog, Twitter, contains more than 199 million daily active accounts in the first quarter 2021 [14].

### A.2.2 Data Collection from Social Media Users

Data collection is normally considered as the first step of the development processes of a predictive model. Within the context of research from personal data like social media, data collection can be considered as one of the hardest steps, due to ethical issues, transparent data processing, and data privacy [151]. This step begins with how to collect users possibly associated with mental disorders, and then annotate those users.

A dataset can be obtained by two approaches: (1) to directly collect from subjects and (2) to directly aggregate from a social media site. This depends on the purpose of each study and the target platform.

#### A.2.2.1 Direct Collection from Subjects

In medical and social sciences, human subject research is a systematic approach to observe and analyse humans agreeing to take part in a study [208, 209]. Choudhury et al. (2013) adapted the human subject research method to observe and collect data from social network users suffering from mental disorder symptoms [24]. They recruited participants and used a questionnaire to screen them for mental disorders. Eichstaedt et al. (2018) asked patients to participate in their study and used medical records to classify the patients as being depressed [28]. After screening, both studies asked participants for access to their social media data to collect social network profiles.

#### A.2.2.2 Aggregating Data Extracted from Public Posts

Another method is searching target users using keywords. This approach is considered an easy and cost-time effective approach to collect users associated with mental disorders from social media platforms, compared to the previous method. In 2014, Coppersmith et al. introduced an automatic data collection method to search target users with mental health disorders [22]. They used regular expressions or keywords to search for a set of target tweets mentioning the diagnosis of mental health diseases. The returned tweets were investigated and verified as genuine tweets, which revealed mental health diagnosis of the tweet owners. This approach is normally performed in *Twitter*. After receiving a set of target posts, verification and annotation processes will be further conducted.

### A.2.3 User Verification and Annotation

After receiving a set of target users or posts, every user and every owner of the returned posts from a profile-based social media platform needs to be annotated. This step is an important one, because correctly labelled users can help to distinguish between user groups and to analyse differences between them. There are two main approaches to user annotation. The first method is screening users with a questionnaire and another is manual annotation.

#### A.2.3.1 Screening Users with a Questionnaire

A traditional and medical method to screen patients for mental health disorders is the use of a self-report questionnaire. This method has been applied to screen online users with depression [24]. Common questionnaires used to screen users with depression on include the *CES-D* questionnaire as explained in *section A.1.2* and *Patient Health Questionnaire-9 (PHQ-9)*.

#### A.2.3.2 Manual Annotation

Another approach to annotating users with mental disorders is manual annotation. This method is usually associated with the approach to searching users by keywords (see *section A.2.2.2*). Researchers investigate every returned message from a social network platform and manually annotate the tweets.

Klein et al. (2018) provided detailed processes of collecting data for health-related events on social media [210]. Figure A-5 depicts the workflow of annotating users with depression. First, a set of regular expressions is created to search messages mentioning self-expressions or self-declaration related to mental disorder diagnosis via a search API. It returns a set of matching messages, which must be manually verified to identify genuine messages exactly mentioning diagnosis. The profiles of verified users who disclosed their mental illness diagnoses are annotated as *positive*. Public pages publishing mental health information are removed or annotated as *negative*. Finally, we receive a target group or users disclosing their mental illness diagnose.

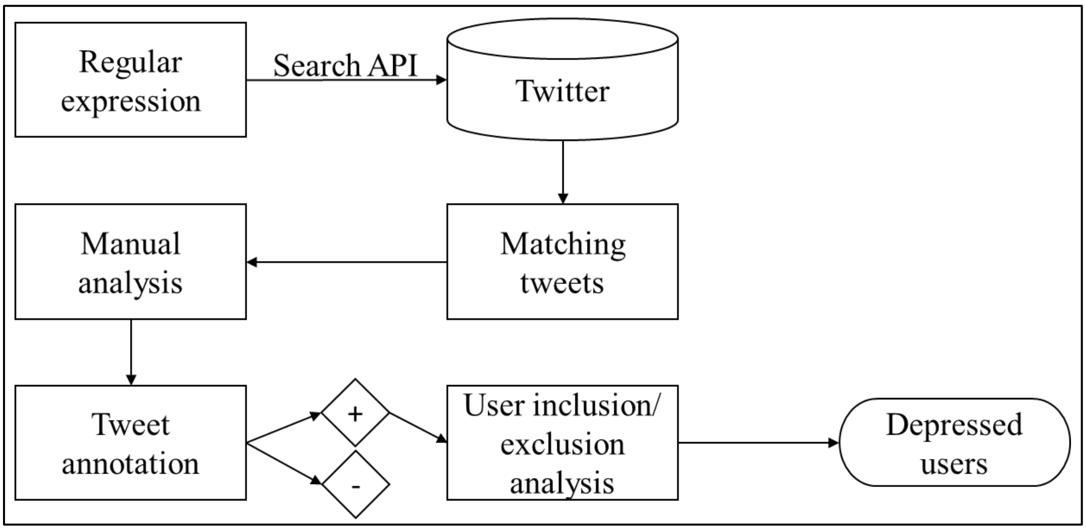


*Figure A-5 The workflow for mining profile-based social media users with depression*

# A.3 Data Collection from Social Media Platforms

There are different methods for collecting profile-based social media data depending on platforms and policies. This section focuses on common and popular social media platforms in which users can update statuses on their timelines and have interactions with friends. In particular, we will focus on data collection from *Facebook* and *Twitter*.

## A.3.1 Facebook

Facebook allows a developer or a company to use their services such as artificial intelligence, gaming, virtual reality, and business tools. One service that researchers are interested and consider valuable for research is called "*Facebook Login*"[6]. An

[6] https://developers.facebook.com/docs/facebook-login/overview

API is provided to develop apps that allow users to log in using their Facebook account and to provide permissions to access their profiles.

To use this function, the developer needs to provide app descriptions and how the app uses users' data in order to get approval from Facebook before releasing the app for public use. There are two approval states (i) app review and (ii) business verification or individual verification[7].

*App review* is a general review process which an app must pass before being published. Apps that request to access Facebook name, an email, and a current profile picture may not go through this process. One permission that is valuable for observing user's behaviours is "*user posts*", which must pass the review process and get approved.

*Business verification or individual verification* is another process to allow Facebook to verify a business or an individual developer before provide access to sensitive user's data, such as user posts, friend lists, and location check ins.

Some permissions are available at both business and individual verification levels, but some are available only at the business verification level. For example, the user post permission is only available with business verification. This means that any app requesting access to user's posts needs to pass both *app review* and *business verification* processes.

[7] https://developers.facebook.com/docs/facebook-login/review

Similar to Twitter search API, Facebook provides a new channel for accessing public Facebook data for researchers. However, the access allows only for public content across Facebook Pages and Groups, as well as verified profiles and public Instagram accounts. In this way, we cannot collect profiles of target users for our research purposes. Currently, Facebook focuses on misinformation, elections, COVID-19, racial justice, well-being only. Other categories listed cannot be accessed using these services[8].

## A.3.2 Twitter

Twitter provides many API channels ranging from publishing tweets to advertising. The most useful APIs that researchers can use to collect user's data are *search tweets* API[9] and *get timelines* API[10].

The *Search tweet* API allows developers to search or query a set of matching tweets using keywords and operators. This API can retrieve tweets since the first tweets in March 2006. Twitter offers three different types of search APIs:

1. *Standard* API provides a sampling of recent Tweets published in the past 7 days.
2. *Premium* API has free and paid access to either the last 30 days of tweets or full access to tweets since 2006.

[8] https://help.crowdtangle.com/en/articles/4302208-crowdtangle-for-academics-and-researchers

[9] https://developer.twitter.com/en/docs/tweets/search/overview

[10] https://developer.twitter.com/en/docs/tweets/timelines/overview

3. *Enterprise* API is a paid version, which can access to both the last 30 days of tweets and full access to tweets since 2006.

The difference between *premium* and *enterprise* APIs is that the former has more restrictions. For example, the enterprise API allows a single query of up to 2,048 characters per request, while the premium API can use up to 1,024 characters. Enterprise API retrieves up to 500 tweets per request and premium API returns the maximum of 100 tweets.

Twitter is releasing the new version of Twitter API v2, which the above rules may be changed[11]. This API also includes "academic research product track", which allows researchers to access Twitter's real-time and full historical public data with no cost[12].

*Get tweet timeline* API is an endpoint to retrieve tweets on the timeline of a user. This API provides up to 3,200 recent tweets on the timeline.

## A.4 Natural Language Processing (NLP)

***Natural language processing (NLP)*** is the field of studying and understanding human languages using computers. NLP studies speech recognition, understanding, machine translation, and generation. It is a multidisciplinary approach consisting of linguistics, computer science, cognitive science, and artificial intelligence [211]. This

[11] https://developer.twitter.com/en/docs/twitter-api/early-access

[12] https://developer.twitter.com/en/products/twitter-api/academic-research

study focuses on natural language understanding to investigate textual content published on social network profiles. The following section will explain descriptions of text processing to process written languages.

***Text processing*** is used to interpret written languages in an appropriate format understandable by a machine. It includes from a simple step like separating given text into smaller parts to a complicated approach. This section explains segmentation, tokenisation, and word embedding.

## A.4.1 Segmentation

***Segmentation*** is the process of separating a given text into smaller segments [212]. It includes two types: one for *sentence segmentation*, and another for *word segmentation*. Sentence segmentation separates a given text into several sentences. It can use question mark, comma, or conjunction words to separate the sentences. This example shows how sentence segmentation segments an English text:

*"I am not feeling well these days, since I have massive workloads 😢 😢 😢."*

This can be segmented into 2 sentences: one for *"I am not feeling well these days"*, and another for *"since I have massive workloads 😢 😢 😢"*. It uses a comma to separate the sentences.

Word segmentation is used to separate a given text into smallest parts [212]. The simplest way of performing word tokenisation is using white spaces, commas, and question marks. From the above example, a tokeniser extracts the text as:

*"I" "am" "not" "feeling" "well" "these" "days" "," "since" "I" "have" "massive" "workloads" "😢" "😢" "😢" "."*

### A.4.2 Linguistic Inquiry and Word Count (LIWC)

***Linguistic Inquiry and Word Count (LIWC)*** is a text analysis program used to extract relevant psychological meanings and linguistic styles from text [213]. LIWC is mainly used in the health analytics and mental health detection from social media data. This tool provides up to 93 dimensions from a given text. The psychologically meaningful categories and function words include, for instance, *affect words* (positive emotion, negative emotion, and sadness), *parts of speech* (1st personal singular pronoun, impersonal pronouns, negations, and regular verbs), *personal concerns* (work, death, and money), and *punctuation* (commas, question marks, and parentheses). For instance, Table A-4 shows the example of extracted dimensions from the sentence in the section A.4.1 by LIWC.

*Table A-4 Extracted dimensions from LIWC*

| **Dimensions** | **Extracted values** |
|---|---|
| 1st personal singular pronoun (*I*) | 16.67 |
| Positive emotion | 8.33 |
| Negative emotion | 0.00 |
| Negations | 8.33 |
| Cognitive Processes (cause) | 8.33 |
| Time Orientation - Present focus | 16.67 |

As can be seen, the extracted values of each category are computed from the percentage of words in the sentence. For example, the word “I” appears twice, and the sentence has twelve words. Therefore, 1st personal singular pronoun is 16.67% (2/12).

### A.4.3 Word Embedding

***Word embedding*** is another text processing technique to transform words or vocabulary of a document into vectors. The main idea of word embedding is to determine word similarity. In other words, a pair of words that frequently co-occur in the same context tends to share similar meanings. *word2vec* and *GloVe* are common techniques used in deep learning.

***Global Vectors (GloVe)*** is an unsupervised learning method for constructing vector space representations of words (Pennington et al., 2014) [214]. GloVe is based on a count-based method, which computes the aggregated distributions of word co-occurrence from a given corpus and performs dimensionality reduction.

To construct GloVe word representation, first a matrix of word-word co-occurrence counts $X$ is created to count the number of times that word $i$ occurs in the context of word $j$. Each cell in the matrix $X$ is represented as $X_{ij}$. Then the probability that the word $i$ appears in the context of word $j$ is computed as $P_{ij} = P(j|i) = \frac{X_{ij}}{X_i}$, where $X_i$ is the number of times that any word occurs in the context of word $i$. Finally, GloVe word representation is given by:

$$J = \sum_{i,j=1}^{V} f(X_{ij})\left(w_i^{\mathrm{T}}\tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$$

Where $V$ denotes the size of the vocabulary, and $f\left(X_{ij}\right)$ is a weighting function. The variables $w_i$ and $\widetilde{w}_j$ are word vectors, while the parameters $b_i$ and $\tilde{b}_j$ denote biases for $w_i$ and $\widetilde{w}_j$, respectively. In the case of rare co-occurrences, one class of functions can be parameterized as:

$$f(x) = \begin{cases} (x/x_{max})^{\alpha} & if\ x < x_{max} \\ 1 & otherwise \end{cases}$$

### A.4.4 Anaphora Resolution

***Anaphora resolution*** is a linguistic method that determines which previously mentioned person is the subject of a subsequent statement [215]. It is the problem of resolving that a reference or an *anaphor* refers to an earlier or later entity or an *antecedent*. For example, anaphora resolution of the message shown below:

"*My friend got diagnosed with depression. He was suffering from his exam failure.*"

This can determine "He" is the anaphor and "My friend" is the antecedent.

## A.5 Machine Learning

***Machine learning*** is a method that allows a computer to learn from a dataset and uses this experience to make a decision [216]. In 1997, Mitchell [217] defined the term "*learning*" as follows "A computer program is said to learn from experience $E$ with respect to some class of tasks $T$ and performance measure $P$, if its performance at tasks in $T$, as measured by $P$, improves with experience $E$." In other words, it means that a computer programme is given a task and its performance can improve with experience.

This section begins with important and necessary terminologies in machine learning. This can help to understand basic concepts in this study.

- An *instance* or an *input* $x$ means a specific object or an observed event. The instance consists of at a least one-dimensional feature vector $x = (x_1, x_2, \dots, x_D) \in \mathbb{R}^D$, where $D$ represents the number of dimensions.

- The term "*feature*" denotes a set of observations that is relevant to the modelling problem, typically represented numerically [218]. Each dimension is often called a feature [219].

- A *label* or an *output* $y$ is a real value associated with an instance $x$. Labels can be either continuous values $\mathbb{R}$ or a finite set of values, e.g., {non-depressed, depressed}. From the example, these categorical values are often called classes and can be represented as $y \in \{0,1\}$, where 0 denotes non-depressed and 1 presents depressed. Two classes can be called binary labels. Some problems can have more than two classes and represented as $y \in \{0, \dots, C\}$, where C denotes the number of classes [219].

- A *training sample* consists of a set of instances $X = \{x_1, x_2, \dots, x_n\}$, where $n$ represents the number of samples or observed objects. Recall that, each instance $x_i$ consists of $D$-dimensional feature vectors [219]. So, the training sample with $n$-instances and $D$-dimensions can be represented as X= $\{(x_{11}, x_{12}, \dots, x_{1D}), (x_{21}, x_{22}, \dots, x_{2D}), \dots, (x_{n1}, x_{n2}, \dots, x_{nD})\}$.

- *Labelled data* is pairs of *(instance, label)*, while *unlabelled data* contains instances alone [219].

## A.5.1 Machine Learning Algorithms

Machine learning algorithms can be broadly divided into two main categories: one for *supervised learning* and another for *unsupervised learning*. Another learning paradigm is *semi-supervised learning*.

### A.5.1.1 Supervised Machine Learning

***Supervised machine learning*** is an algorithm that can learn from a set of inputs supervised by a set of labels paired with the inputs. In other words, it refers to the method followed by a machine able to learn a set of patterns from provided data with the pair of their labels and to make a prediction based on the learnt patterns [216]. The training sample $\{(x_i, y_i)\}_{i=1}^{n}$ contains a set of pairs between instances $X = \{x_1, x_2, \dots, x_i\}$ and labels $Y = \{y_1, y_2, \dots, y_i\}$. The learning algorithm can then learn from the training sample and try to make a prediction $\hat{y}$ by:

$$p(y|x) = \frac{p(x,y)}{\sum_{\hat{y}} p(x,\hat{y})}$$

Examples of supervised learning algorithms include *regression*, *support vector machine (SVM)*, *naïve Bayes*, and *decision trees*.

### A.5.1.2 Unsupervised Machine Learning

***Unsupervised machine learning*** is a method that captures unknown patterns or features in a dataset and then produces the most suitable representation associated with the dataset [216]. Unsupervised machine learning is the chain rule of probability from inputs $X = \{x_1, x_2, \dots, x_i\}$. This can be represented as:

$$p(x) = \prod_{i=1}^{n} p(x_i | x_1, x_2, \dots, x_{i-1})$$

To distinguish between supervised and unsupervised machine learning, the former requires both a feature and a label, while the latter may require only a feature. Unsupervised learning algorithms include *principal components analysis*, and *k-means clustering.*

### A.5.1.3 Semi-supervised learning

***Semi-supervised learning*** is the combination of the *supervised* and the *unsupervised* learning. Semi-supervised classification is an extension to the supervised learning, which the training data contains both labelled and unlabelled instances. This reflects partially labelled data learning. One of well-known semi-supervised learning algorithms is *multiple instance learning (MIL)*.

## A.5.2 Multiple Instance Learning (MIL)

***Multiple instance learning (MIL)*** is a weakly supervised learning algorithm first proposed by Keeler, Rumelhart, and Leow (1991) [220]. As mentioned above, supervised learning requires instances and labels associated with the instances to learn during the training process. The main advantage of MIL is that instead of requiring a single instance, MIL can learn from instances bags $\{X_1, X_2, \cdots, X_i\}$ and labelled bag $y_1, y_2, \ldots, y_i$, where each $X_i$ contains instances $\{x_{i1}, x_{i2}, \ldots, x_{in}\}, x_{in} \in X$. Each instance can be independent. In such way, each instance can have its own individual label $\{y_{i1}, y_{i2}, \cdots, y_{in}\}$, where each $y_{in} \in \{0, 1\}$. It is assumed that each $y_{in}$ is unknown during the training process. From these assumptions, a MIL classifier can predict a label $Y$ for a given bag $X$ by:

$$y_i = \begin{cases} 1, & if\ \exists \mathrm{y}_{in} = 1 \\ 0, & otherwise \end{cases}$$

MIL can then be trained as either an instance classifier $f(X): X \to Y$ or a bag classifier $F(X): X^i \to Y$. From the assumption above, MIL can provide an extreme result $y_i = 1$ in the case of having an instance $x_{in}$ predicted as a positive label $y_{in} = 1$.

The relaxation of the MIL assumption can compute using distributions of instances. Each bag label can be computed from all probability distributions of instances $y_i = P(X_i)$, where $X_i = \{x_{i1}, x_{i2}, \dots, x_{in}\}$.

## A.6 Deep Learning

Machine learning techniques have been successfully developed, improved and applied in many studies. However, classical machine learning algorithms may need handed feature engineering, which is time expensive. Sometimes they cannot perform well on high-dimensional datasets. A new type of machine learning algorithm implemented to overcome this problem is ***deep learning*** or ***deep neural networks*** [216].

***Deep learning*** has a long history of implementations, usually referred to with different names. In 1958, a neural network trained with one or two hidden layers was introduced [221]. The name "*deep learning*" is given, due to the depth of many connected layers in a model. The term of *neural* is borrowed from neuroscience. Each next hidden layer of a neural network is connected and represented hidden vector values sent to the next layer. This is assumed to be similar to how a brain neuron works, where each cell receives signal from a previous cell and sends it to the next cell.

Deep learning is useful to simplify representations from raw data, which may contain complex concepts such as high-level and abstract features. Figure A-6 shows the illustration of a deep learning model extracting components from pixels of pictures/inputs and then map extractable representations to outputs/labels. Different parts of the deep learning model represent edges, corners, contours, and objects parts extracted from the picture. Those components are then mapped to the output “*person*”.

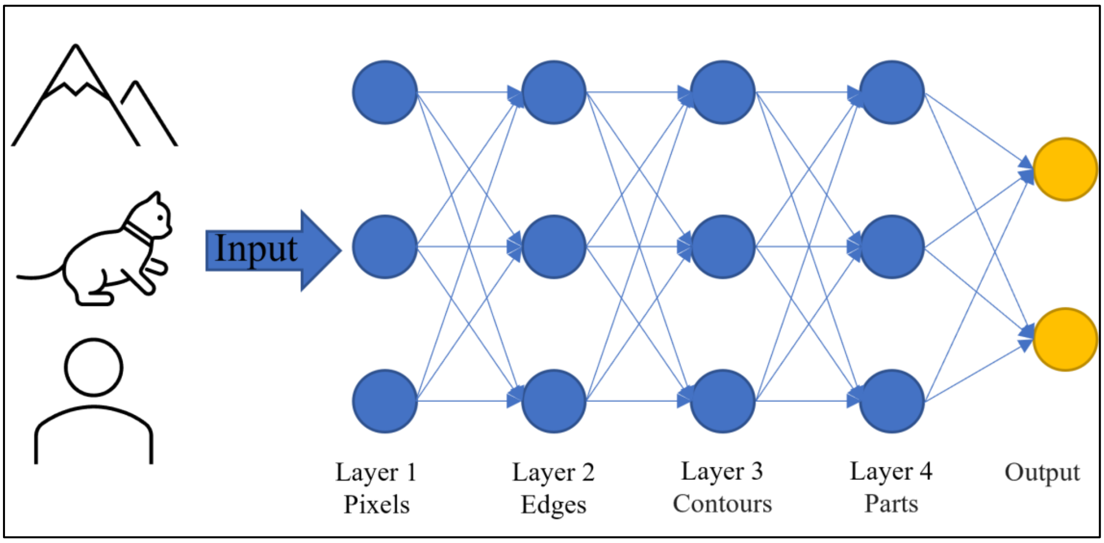


*Figure A-6 A deep learning model to extract high-level and abstract features from pictures.*

This section introduces the variety of deep learning techniques intensively and successfully used to build a prediction model. A family of neural architectures includes the basic layer of deep learning called ***multiple layer perceptron (MLP)***. Other type of layers used in computer vision is ***convolutional neural network (CNN)*** and in natural language processing is ***recurrent neural network (RNN)***. The improved version of RNN is ***long short-term memory (LSTM)***.

### A.6.1 Multiple Layer Perceptron (MLP)

***Multiple layer perceptron (MLP)*** or ***deep feedforward network*** is a basic and essential deep learning layer. It computes the approximation of a function $f^*$ via parameters $\theta$ to provide an output associated with the learnt function and a provided input. In other words, a classifier with the MLP is a function $y = f(x; \theta)$ which maps an input $x$ to an output $y$ and computes the best approximation value of the parameters $\theta$ suitable for the output $y$ [216].

Developing a deep learning model requires the design of an *output unit* to compute the form of a desired output and a *cost function* to measure the differences between provided outputs and predicted results in the network. An *activation function* is chosen to compute hidden values in each layer. The overall architecture of the model needs to be taken into account in order to design how many layers the network should have, how each layer is connected to each other, and how many units should have in each layer [216].

The designed deep learning is then trained to produce an approximated output $\hat{y}$ from the provided input $x$. This results in the flow of information of $x$ throughout the network. This is called *forward propagation*. During training, the cost (an error result occurs between a provided output and a predicted value) is computed and then is flowed backward through the network to improve the predicted result. This process is called the *back-propagation* algorithm or *backprop*.

Figure A-6 shows how a neural network works. However, deep learning consists of various types of layers. Each layer of a neural network model can be replaced by *CNN*, *RNN*, and *LSTM*.

### A.6.2 Convolutional Neural Network (CNN)

***Convolutional network*** or ***Convolutional neural network (CNN)*** [216] was introduced by LeCun [222, 223] and first implemented by Fukushima [224]. CNN is one among several well-known neural networks introduced to process data like a grid topology.

CNN normally performs two operations. The first step is called a *convolutional layer/operation*. CNN creates a *kernel* or *filter* with the size of width and height. The *kernel* is shifted spatially along with the size of an *input*. After shifting the kernel along all the area of the input, the convolution operation provides an *output,* sometimes called a *feature map*. The size of the *kernel* is assumed to be smaller than the size of the input, which helps the *kernel* to provide a set of learnable features.

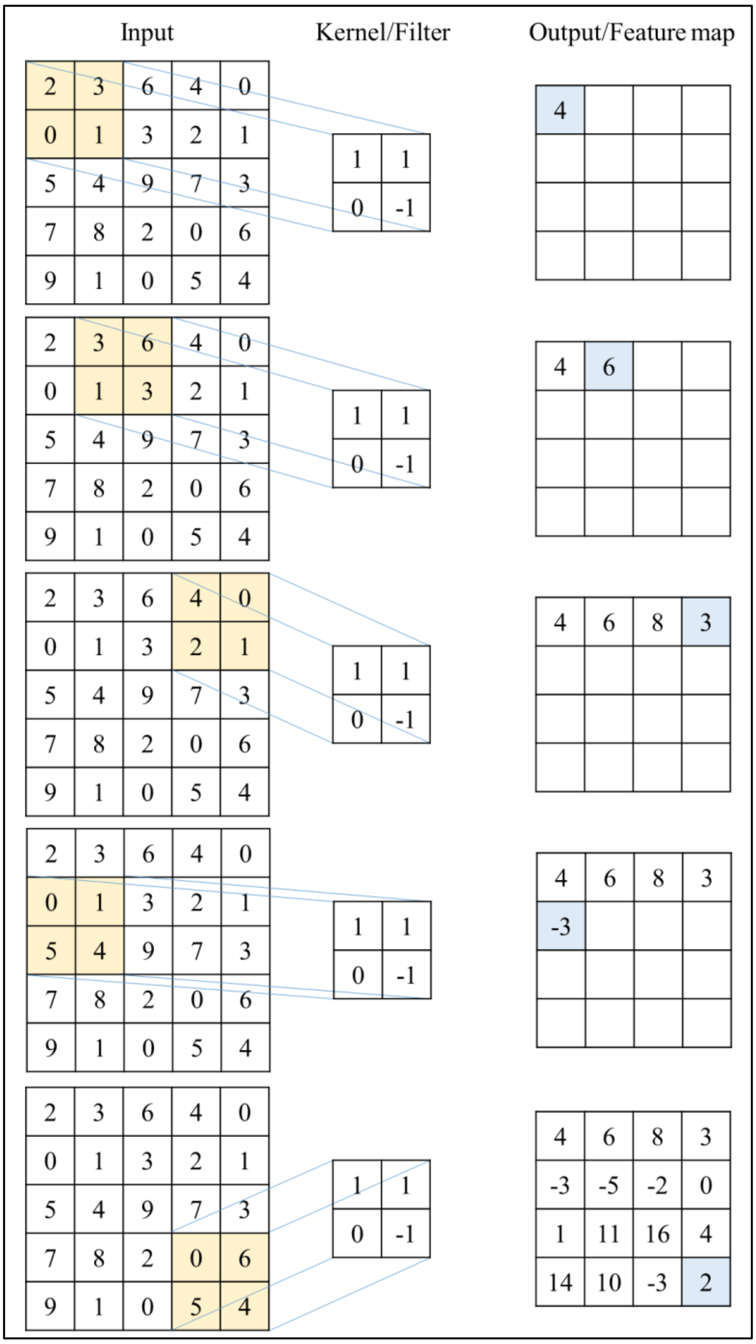


*Figure A-7 The convolutional operations. The kernel of size 2×2 with a stride 1 performs dot products with local regions of the input. The kernel is shifted along both width and height of the input size 5×5. The yellow area represents the local regions that the kernel is hovering, and the blue area is the computed results received from dot products between the kernel and the hovered area.*

Figure A-7 shows how the convolutional operation works. The kernel of size 2×2 is created and initially random its values. The kernel is applied with a stride of 1, which means it is shifted spatially with every one-cell from left to right and from the top to the bottom along the input size. In every shift of the kernel, it performs dot products with local regions of the input where the kernel is hovering over. As shown in the equation below, the kernel and the local region of the input performs matrix multiplication. This returns 4 in the first cell of the feature map/output in Figure A-7.

$$\begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix} = (2 \times 1) + (3 \times 1) + (0 \times 0) + (1 \times -1) = 4$$

Receiving the feature map, a *pooling layer/operation* is applied to reduce the spatial size of an input and preserve the originality of the input. The reduction also helps to reduce the number of parameters in the networks and avoid overfitting. The pooling layer consists of two operations. The *MAX pooling* operation provides the maximum value in the local region where the filter is hovering over. The *AVERAGE pooling* operation computes the average value in the region in which the filter is hovering over. The filter is shifted all over the input area and then a pooling output is received, as shown in the Figure A-8 for MAX pooling operation and Figure A-9 for AVERAGE pooling operation.

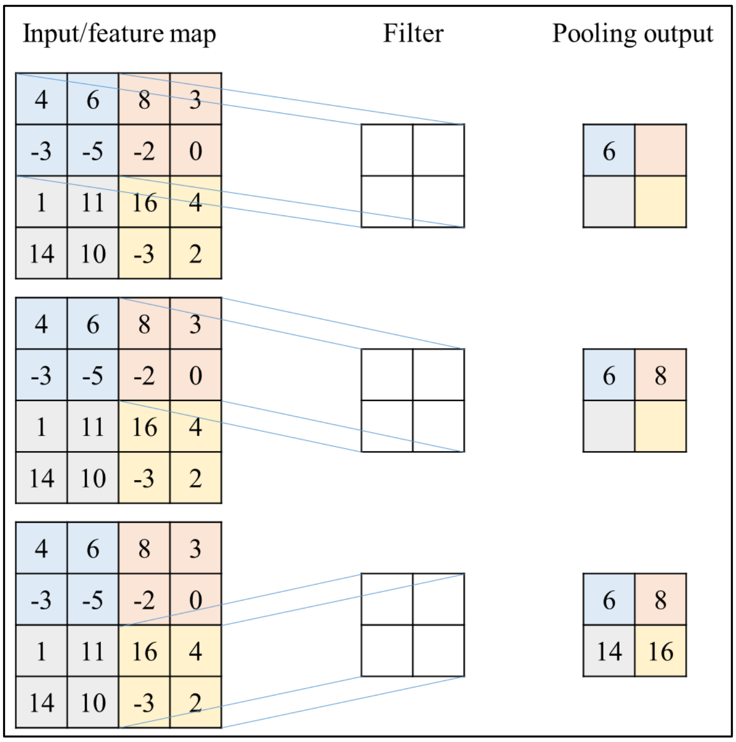


*Figure A-8 The MAX pooling operation over an input. The filter is created with the size of 2X2 and a stride 2. The different colour indicates the different areas where the filter is shifted over.*

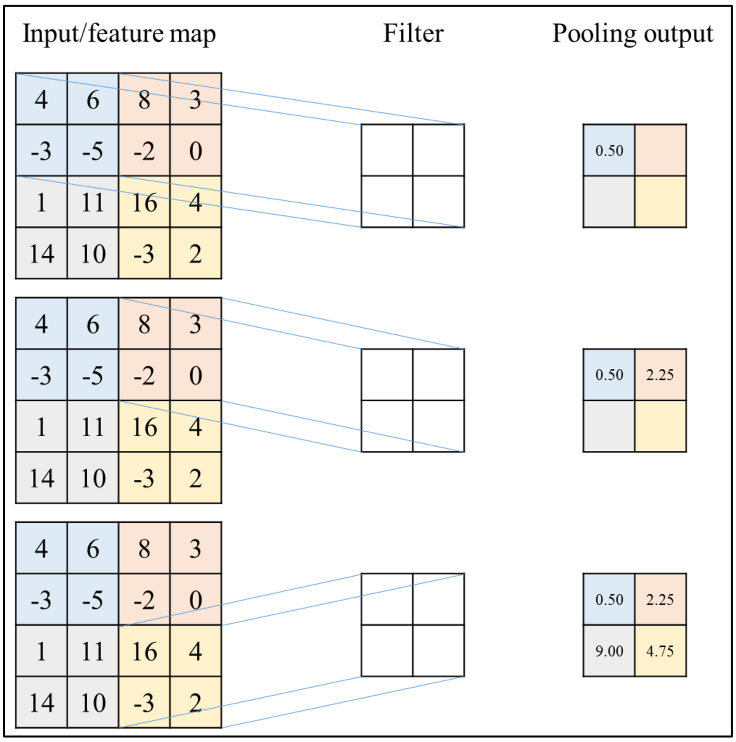


*Figure A-9 The AVERAGE pooling operation over an input. The filter is created with the size of 2X2 and a stride 2. The different colour indicates the different areas where the filter is shifted over.*

### A.6.3 Recurrent Neural Network (RNN)

***Recurrent Neural Network (RNN)*** [216] was implemented by Rumelhart et al. [221]. The main purpose of RNN is to process the sequence of data. RNN provides the output members of the network by learning from the outputs of the previous state.

$$h^{(t)} = f\left(h^{(t-1)}, x^{(t)}; \theta\right),$$

where $f$ is a function, $h^{(t)}$ is the state of a system, $x^{(t)}$ is an input at timestep $t$, and $\theta$ is a value used to parametrise the function $f$. The above equation can be represented as the unfolded equation with a function $g^{(t)}$:

$$h^{(t)} = g^{(t)}(x^{(t)}, x^{(t-1)}, x^{(t-2)}, \cdots, x^{(2)}, x^{(1)})$$

When taking the RNN into a deep learning structure, the above equation is represented with forward propagation as:

$$a^{(t)} = b + Wh^{(t-1)} + Ux^{(t)},$$

$$h^{(t)} = \tanh\left(a^{(t)}\right),$$

$$o^{(t)} = c + Vh^{(t)},$$

$$\hat{y}^{(t)} = softmax\left(o^{(t)}\right),$$

where $b$ and $c$ are bias vectors, $U$ is the weight matrix for input-to-hidden, $V$ is the weight matrix for hidden-to-output, and $W$ is the other matrix for hidden-to-hidden. For each timestep $t$, the parameters $a$, $h$, $o$, and $\hat{y}$ are the activation, the hidden representation, the output, and the target, respectively. The equation can be represented as a computational graph, as shown in Figure A-10.

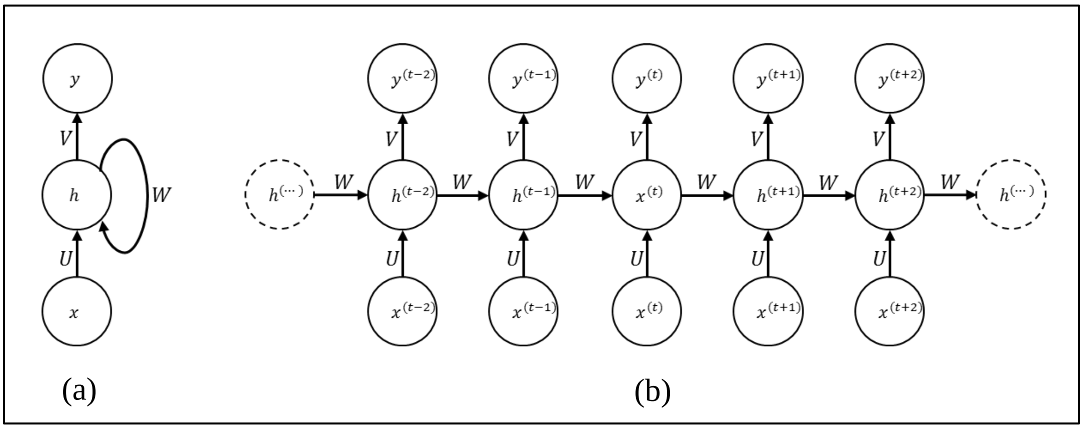


*Figure A-10 The computational graph of a RNN. (a) The RNN structure with recurrent connections. (b) The time-unfolded computational graph of the RNN.*

The RNN structure has few drawbacks. Its computation is slow and cannot properly learn long-tern dependencies of an input. One way of dealing with these issues is to design a model that partitions the long-tern dependencies into small parts and processes with multiple time scales. In this way, the model can transfer information from the distant past to the present.

### A.6.4 Long Short-Term Memory (LSTM)

*Long short-term memory (LSTM)* was introduced by Hochreiter and Schmidhuber in 1997 to eliminate the above problems [225]. Instead of a fixed loop, LSTM decides the weight on its self-loop depending on another previous hidden unit on the context of its inputs. In this way, LSTM is more flexible and can changes the time scale dynamically. A LSTM cell consists of *a forget gate* $f^{(t)}$, *an input gate* $i^{(t)}$, *an output gate* $o^{(t)}$, *new cell content* $\tilde{C}^{(t)}$, *a cell state* $C^{(t)}$, and *a hidden state* $h^{(t)}$. The internal LSTM cell is represented in Figure A-11. LSTM equation can be represented as:

$$f^{(t)} = \sigma(W_f h^{(t-1)} + U_f x^{(t)} + b_f)$$

$$i^{(t)} = \sigma(W_i h^{(t-1)} + U_i x^{(t)} + b_i)$$

$$o^{(t)} = \sigma(W_o h^{(t-1)} + U_o x^{(t)} + b_o)$$

$$\tilde{C}^{(t)} = tanh(W_c h^{(t-1)} + U_c x^{(t)} + b_c)$$

$$C = f^{(t)} \circ c^{(t-1)} + i^{(t)} \circ \tilde{C}^{(t)}$$

Where $U$, $W$, and $b$ are the *input weights*, *recurrent weights*, and *biases*, respectively, of each above equation. The parameters $x^{(t)}$ denotes the current input vector and $h^{(t)}$

represents the current hidden vector. After computing those gates, hidden outputs can be received as following:

$$h^{(t)} = o^{(t)} \circ C^{(t)}.$$

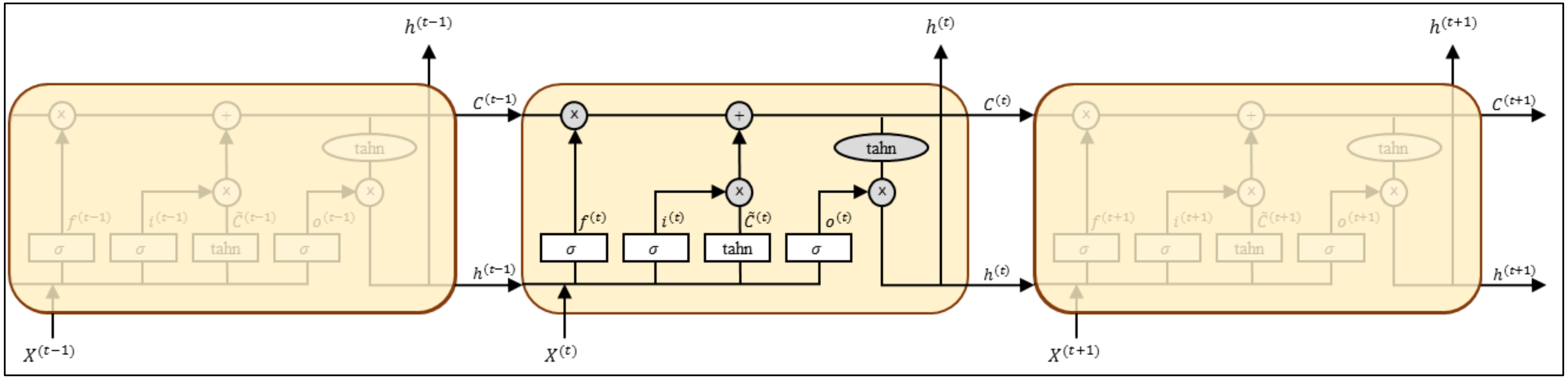


*Figure A-11 The internal LSTM cells*

# A.7 Evaluating ML Models

This section explains evaluation methods to measure the performance of a model. It begins addressing how to split a dataset into small subsets. Then evaluation matrices will be described. Finally, we provide details of model selection.

## A.7.1 Cross Validation

***Cross-validation*** is a technique to split a dataset into different training and test sets [226]. In machine learning settings, a *training set* is the dataset used to develop or train a machine learning model. A *test set* is separated to evaluate the model, and the test set is unseen during the training step. Cross-validation randomly partitions a dataset into *n* equal subsets and proceeds to iterate *n* times, with each subset used for validation or testing exactly once, while the remaining *n*–1 subsets are used as the training data. For instance, Figure A-12 illustrates 10-fold cross validation, which splits a dataset into 10 equal subsets. In each iteration, the *n–1* training subsets are used to train a model and then the unique test set is used to evaluate performance of the model.

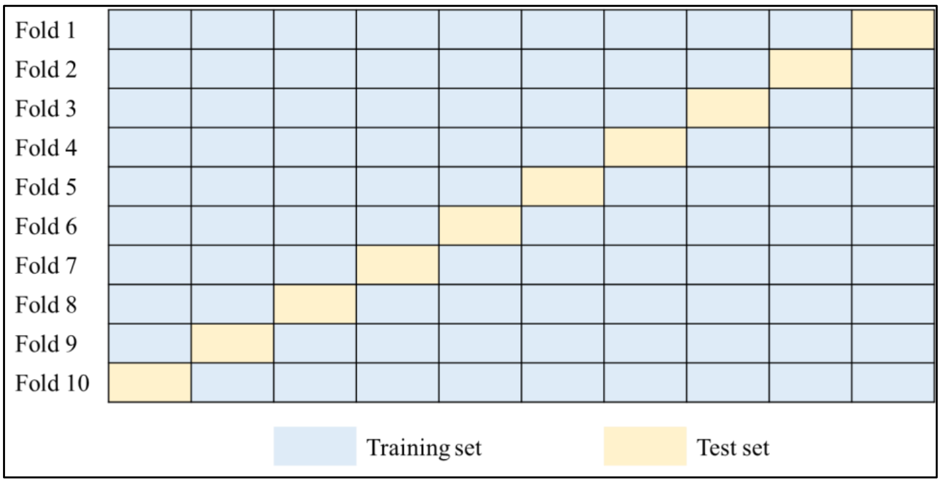


*Figure A-12 The 10-fold cross validation.*

## A.7.2 Classification Evaluation

***Classification evaluation*** is an approach to measure performance of a classifier. Classification problems include *binary classification* and *multiclass classification*. Binary classification is the task of classifying samples into one of only two different classes, while multiclass classification means a method of determining instances as one of more than two different classes. For example, classifying users with and without depression is binary classification, and classifying a message into one of negative, positive and neutral classes reflects the multiclass classification problem. This study focuses on a binary classification problem. Results of binary classification can be represented in a confusion matrix (see Table A-5).

*Table A-5 Confusion matrix for binary classification*

| **Sample classes** | **Classified results** | |
|---|---|---|
| | **Positive** | **Negative** |
| **Positive** | True Positive (TP) | False Negative (FN) |
| **Negative** | False Positive (FP) | True Negative (TN) |

The number of each class represented in the confusion matrix can be used to measure the performance of a classifier. A set of common tools for a binary classification problem consist of accuracy, precision, recall (sensitivity), specificity, f1-score, receiver operating characteristic (ROC), and area under the curve (AUC). These metrics are computed using the following formulas:

#### A.7.2.1 Accuracy

***Accuracy*** takes correctly classified results of both classes and the total number of samples into account.

$$Accuracy = \frac{TP + TN}{\sum Population}$$

#### A.7.2.2 Precision

***Precision*** is the measurement of the proportion of positive results correctly classified to the number of all positive results retrieved from a classifier. In other words, the number of correct instances retrieved from a classifier is divided by the number in the column of positive test results in Table A-5.

$$Precision = \frac{TP}{TP + FP}$$

#### A.7.2.3 Recall

***Recall*** or ***sensitivity*** (as commonly called in the medical field) is the effectiveness of a classifier at identifying samples with a positive class. This reflects a true positive rate or how correctly a classifier provides the fraction of true positive samples. Recall measures the ratio of correctly classified results to the total number of samples in a positive class. In other ways, the number of true positive instances retrieved by a classifier is divided by the total number in the positive row in Table A-5.

$$Recall = \frac{TP}{TP + FN}$$

### A.7.2.4 Specificity

***Specificity*** measures the effectiveness of a predictive model correctly classifying samples with a negative class. The method is used frequently in the medical field. This highlights a true negative rate or how effectively a classifier identifies true negative samples. Specificity takes the ratio of correctly predicted results to the total number of samples with a negative class. From Table A-5, specificity can be measure by dividing the number of true negative instances by the total number in the negative row.

$$Specificity = \frac{TN}{TN + FP}$$

### A.7.2.5 F-score

***F-score*** measures the weights between precision and recall. The F-score is calculated based on a weight function $\beta$. Computing the F-score takes the ratio of precision to recall, which can be represented as:

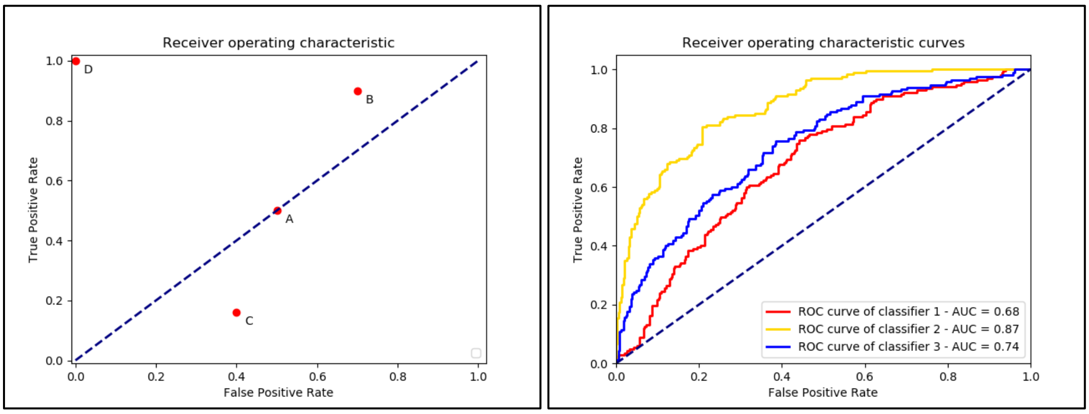
$$F - score = (1 + \beta^2) * \frac{Precision * Recall}{\beta^2 * Precision + Recall}.$$

To equalise between precision and recall, the weight function $\beta$ can be replaced by 1. F1-score is finally represented as:

$$F1 - score = \frac{2 * TP}{2 * TP + FP + FN}.$$

## A.7.3 Receiver Operating Characteristic (ROC)

***Receiver Operating Characteristic (ROC)*** is a graphical and useful tool for measuring and visualising the relationship between the true positive rate and the false positive rate [227]. Figure A-13 shows ROC graphs plotted between 0 and 1 on *x* and *y* axes. Figure A-13 (a) shows four classifiers. Models on the upper left-hand side of the graph are better. The model *D* on the point (0, 1) reflects a perfect classification task, which can provide all classified results without any wrongly classified sample. The model *A* is on the guessing line (the blue dashed line), a sign of randomly guessing results. The model *C* is under the guessing line, meaning that the model performing worse than randomly guessing results. The results of ROC can be used to compare classifiers and select the best classifier. The results of the ROC can be depicted in two different types. A single point is plotted in the ROC space, in case of a classifier providing only a class decision such as *depressed* or *non-depressed*, as shown in Figure A-13 (a). A ROC curve is visualised in case of a classifier producing probabilities or scores for classes, as shown in Figure A-13 (b).



(a)

(b)

*Figure A-13 ROC graphs (a) showing class decision from four classifiers (b) showing ROC curves and AUCs computed from probabilities of three classifiers*

### A.7.4 Area Under the ROC Curve (AUC)

A single scalar value can be computed from the ***area under the ROC curve (AUC)***. The value represents performance of the model underlying the curve [227, 228]. AUCs from different classifiers can be used to compare the performance of models and choose the best classifier. The values of AUC range from 0 to 1. The value is not expected to be lower than 0.5, which goes along with a ROC curve above the randomly guessing line. Figure A-13 (b) illustrates AUCs computed from probabilities of three different classifiers.

### A.7.5 Akaike Information Criterion (AIC)

***Akaike Information Criterion (AIC)*** is a model selection technique to compare how well each model performs. AIC is a commonly used tool for model comparison and model selection [229, 230] that measures the information loss in each model, taking into account the model's complexity as well. AIC is defined as:

$$AIC = -2\,ln(\hat{L}) + 2K + \frac{2K^2 + 2K}{n - K - 1}$$

Where $n$ is the number of samples, and $K$ is the number of parameters or features of a model. $ln(\hat{L})$ denotes the natural logarithm of likelihood [231]. The equation also uses bias adjustment due to a small sample size [232, 233]. Lower AIC values indicate better performance.

## A.8 Summary

This chapter provided the basic background for the development of a predictive model for detecting users with mental disorders. Each mental illness has different characteristics, and some are common. Data collection methods vary depending on the chosen social network platforms and there are differences in labelling samples depending on the sources of collected data from participants. Mining textual data requires a set of text processing tools to explore and understand the meaning of the text, allowing machine learning techniques to distinguish differences among samples. Model evaluation is an important part of the methodology as it allows comparison between data sources and algorithms used to develop the classifiers.

# Appendix B

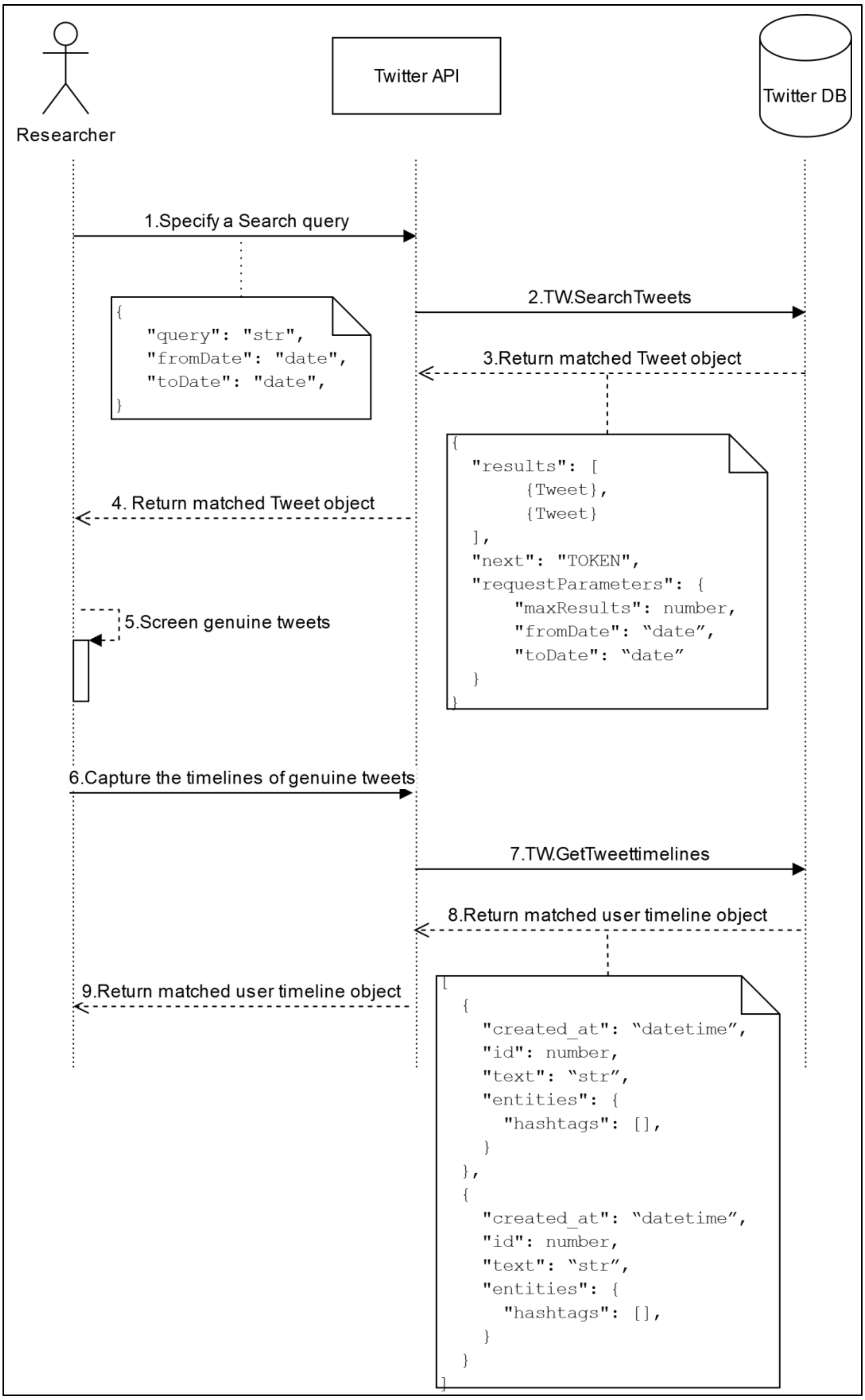
# Data Capture Toolkit for Profile-Based Social Media

Following the explanation of the general data capture process in Chapter 2, this chapter will develop a data capture toolkit for user data on profile-based social media platforms, namely one microblogging platform, especially *Twitter*, and one social network site, *Facebook*.

## B.1 Data Capture from a Microblogging Platform

This section will explain the technical details of data collection from a microblogging platform. Each platform has its own data access channel or provides an API to access data. In this study, we will show the connection between our data collection tool and a Twitter API.

Figure B-1 depicts the technical details of searching a set of genuine messages mentioning the diagnosis of depression and collecting timelines of users. It starts with searching control and target groups. The matched tweets form the searches will be manually screened and followed by gathering timelines of the users screened. The detailed information of data capture processes will be provided below.


Researcher
Twitter API
Twitter DB
1.Specify a Search query
2.TW.SearchTweets
3.Return matched Tweet object
4. Return matched Tweet object
5.Screen genuine tweets
6.Capture the timelines of genuine tweets
7.TW.GetTweettimelines
8.Return matched user timeline object
9.Return matched user timeline object

```
{
   "query": "str",
   "fromDate": "date",
   "toDate": "date",
}
```

```
{
  "results": [
       {Tweet},
       {Tweet}
  ],
  "next": "TOKEN",
  "requestParameters": {
      "maxResults": number,
      "fromDate": "date",
      "toDate": "date"
  }
}
```

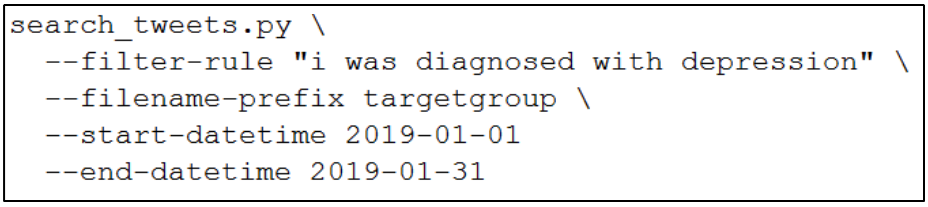
```
[
  {
    "created_at": "datetime",
    "id": number,
    "text": "str",
    "entities": {
      "hashtags": [],
    }
  },
  {
    "created_at": "datetime",
    "id": number,
    "text": "str",
    "entities": {
      "hashtags": [],
    }
  }
]
```

*Figure B-1 The sequence diagram of the technical details of collecting data from a microblogging platform.*

### B.1.1 Target Group Search

This begins with specifying a search query which is “I was diagnosed with depression”. The command shown in Figure B-2 is used to retrieve a set of statuses consisting of “I”, “was”, “diagnosed”, “with”, and “depression” words via a Python Twitter Search API[13], providing and supporting a command-line utility for searching tweets. The API returns a set of matched messages between 1st Jan and 31st Jan 2019. The date range is an arbitrary date range to illustrate the example data collection. The returned results are stored in the JSON file named “targetgroup.json”. JSON (JavaScript Object Notation) is a human-readable and machine data-interchange format, represented as attribute–value pairs and array data types. Figure B-3 shows the example of JSON format.

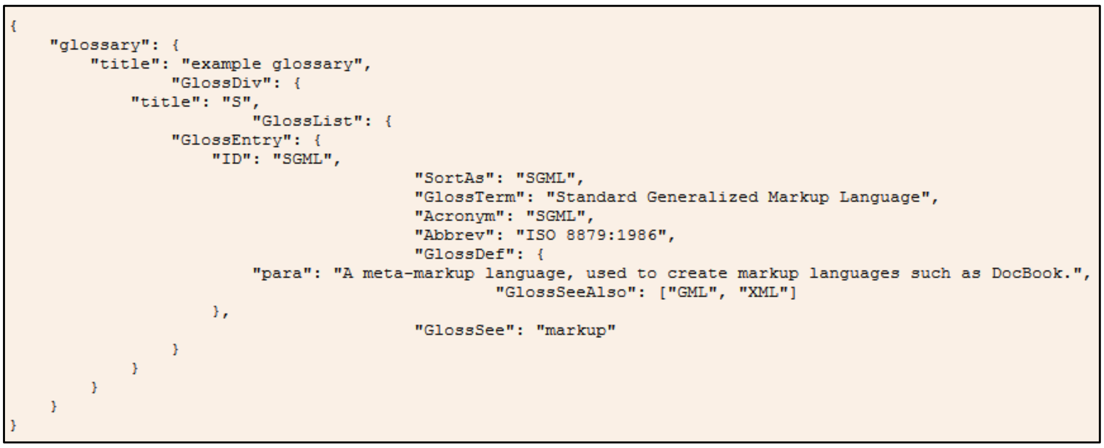
A set of matched statuses are screened for genuine statuses truly revealing the diagnosis of depression. During this process, a human would manually screen for genuine tweets. The set of the messages passing our screening criteria are kept for downloading their timelines.



```
search_tweets.py \
  --filter-rule "i was diagnosed with depression" \
  --filename-prefix targetgroup \
  --start-datetime 2019-01-01
  --end-datetime 2019-01-31
```

*Figure B-2 The example of the command line for collecting target users*

[13] https://github.com/twitterdev/search-tweets-python

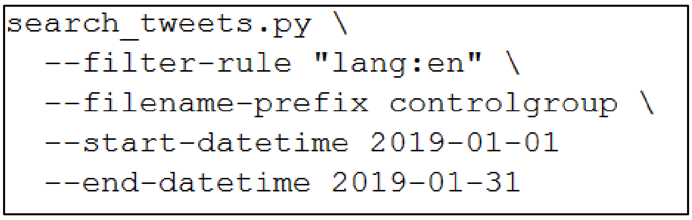


```
{
    "glossary": {
        "title": "example glossary",
                "GlossDiv": {
            "title": "S",
                        "GlossList": {
                "GlossEntry": {
                    "ID": "SGML",
                                        "SortAs": "SGML",
                                        "GlossTerm": "Standard Generalized Markup Language",
                                        "Acronym": "SGML",
                                        "Abbrev": "ISO 8879:1986",
                                        "GlossDef": {
                        "para": "A meta-markup language, used to create markup languages such as DocBook.",
                                                "GlossSeeAlso": ["GML", "XML"]
                    },
                                        "GlossSee": "markup"
                }
            }
        }
    }
}
```

*Figure B-3 The example of JSON format taken from json.org*[14]*.*

## B.1.2 Control Group Search

Similarly, to collect control users, Figure B-4 shows the command for capturing users posting in English between 1st Jan and 31st Jan 2019 and saving results to "controlgroup.json". The date range is just arbitrary. To search the control users, no specific search terms were used. The user IDs of matched tweets are kept for user-timeline collection.



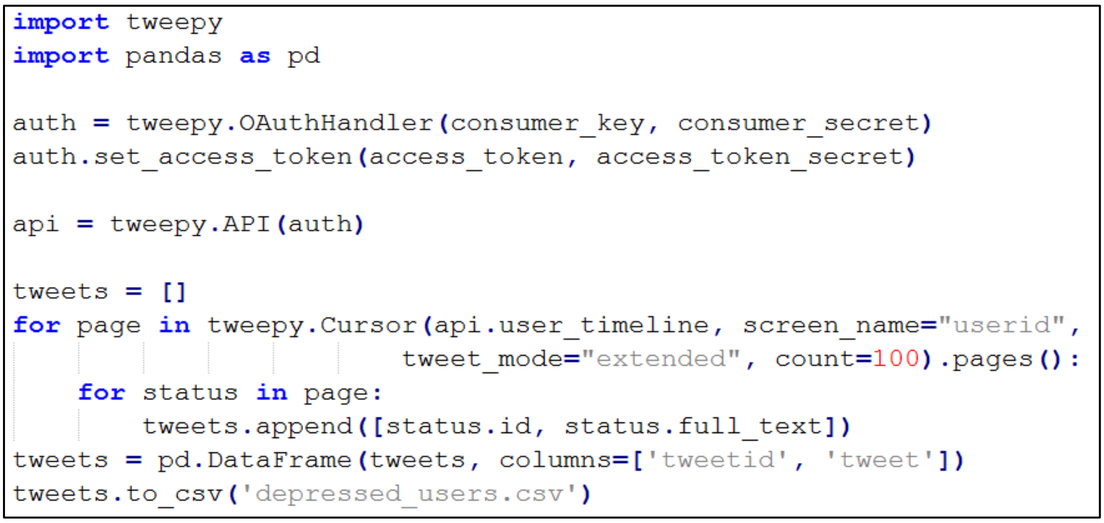
```
search_tweets.py \
  --filter-rule "lang:en" \
  --filename-prefix controlgroup \
  --start-datetime 2019-01-01
  --end-datetime 2019-01-31
```

*Figure B-4 The example of the command line for collecting control users*

[14] https://json.org/example.html

### B.1.3 User Timeline Capture

Tweepy[15], a Python library for accessing the Twitter API, is used to gather timelines of those users. Figure B-5 shows the Tweepy code for collecting the timeline of a user. It connects to a user timeline API returning up to 3200 most recent statuses posted from the specified user[16]. The returned statuses are saved into a comma-separated values (CSV) file, which is a plain text file using a comma to separate values and can be saved in a tabular format.



```
import tweepy
import pandas as pd

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

api = tweepy.API(auth)

tweets = []
for page in tweepy.Cursor(api.user_timeline, screen_name="userid",
                        tweet_mode="extended", count=100).pages():
    for status in page:
        tweets.append([status.id, status.full_text])
tweets = pd.DataFrame(tweets, columns=['tweetid', 'tweet'])
tweets.to_csv('depressed_users.csv')
```

*Figure B-5 The Python code for collecting the timeline of a user.*

[15] https://www.tweepy.org/
[16] http://docs.tweepy.org/en/latest/api.html

## B.2 Data Capture form a Social Network Platform

The difference between a microblogging and a social network platform requires a different mechanism for collecting data. A microblogging platform is relatively public, while a social network platform is more private. This requires a secure approach to connect to a user profile granted permissions. This section will explain the technical details of capturing user data from a social network platform, especially Facebook.

Developing an application on Facebook has four different platforms including iOS, Android, website, and devices (smart TV, camera, and printer)[17]. The development of a Facebook application requires app review[18] and business verification[19] (as explained in *appendix A.3.1*) to verify that the app uses Facebook services in an approved manner.

In this study, our data collection tool is a web-based application due to ease of accessibility from anytime, anywhere, and any device. Our toolkit is separated into a participant section and a research member section.

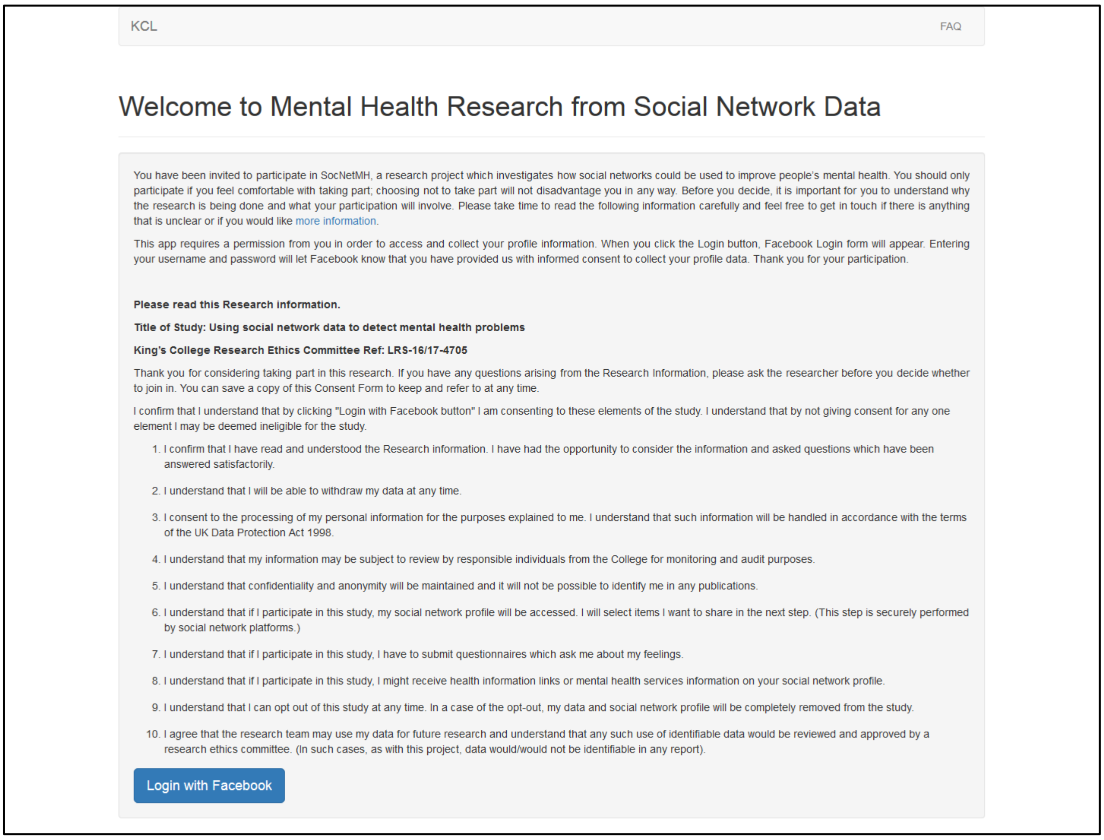
### B.2.1 Data Capture Tool for Participants

This will provide the technical details of the data collection tool for participants. This includes two main modules: one for a login; and another for a user portal.

[17] https://developers.facebook.com/docs/facebook-login/for-devices
[18] https://developers.facebook.com/docs/app-review/
[19] https://developers.facebook.com/docs/apps/business-verification

### B.2.1.1 User Login

User login is a mechanism for participants or users to entry our data collection application and provide permissions for the use of their social network profiles. Figure B-7 illustrates the sequence diagram representing the login process for participants. It begins with a user accessing our application which its index page shows research information with the ethical approval number and informs consent for participating our study. Figure B-6 displays the screenshot of the index page of the application.



Welcome to Mental Health Research from Social Network Data

You have been invited to participate in SocNetMH, a research project which investigates how social networks could be used to improve people's mental health. You should only participate if you feel comfortable with taking part; choosing not to take part will not disadvantage you in any way. Before you decide, it is important for you to understand why the research is being done and what your participation will involve. Please take time to read the following information carefully and feel free to get in touch if there is anything that is unclear or if you would like more information.

This app requires a permission from you in order to access and collect your profile information. When you click the Login button, Facebook Login form will appear. Entering your username and password will let Facebook know that you have provided us with informed consent to collect your profile data. Thank you for your participation.

**Please read this Research information.**

**Title of Study: Using social network data to detect mental health problems**

**King's College Research Ethics Committee Ref: LRS-16/17-4705**

Thank you for considering taking part in this research. If you have any questions arising from the Research Information, please ask the researcher before you decide whether to join in. You can save a copy of this Consent Form to keep and refer to at any time.

I confirm that I understand that by clicking "Login with Facebook button" I am consenting to these elements of the study. I understand that by not giving consent for any one element I may be deemed ineligible for the study.

1. I confirm that I have read and understood the Research information. I have had the opportunity to consider the information and asked questions which have been answered satisfactorily.
2. I understand that I will be able to withdraw my data at any time.
3. I consent to the processing of my personal information for the purposes explained to me. I understand that such information will be handled in accordance with the terms of the UK Data Protection Act 1998.
4. I understand that my information may be subject to review by responsible individuals from the College for monitoring and audit purposes.
5. I understand that confidentiality and anonymity will be maintained and it will not be possible to identify me in any publications.
6. I understand that if I participate in this study, my social network profile will be accessed. I will select items I want to share in the next step. (This step is securely performed by social network platforms.)
7. I understand that if I participate in this study, I have to submit questionnaires which ask me about my feelings.
8. I understand that if I participate in this study, I might receive health information links or mental health services information on your social network profile.
9. I understand that I can opt out of this study at any time. In a case of the opt-out, my data and social network profile will be completely removed from the study.
10. I agree that the research team may use my data for future research and understand that any such use of identifiable data would be reviewed and approved by a research ethics committee. (In such cases, as with this project, data would/would not be identifiable in any report).

*Figure B-6 The screenshot of an index page of our data collection tool to show research information and informed consent.*

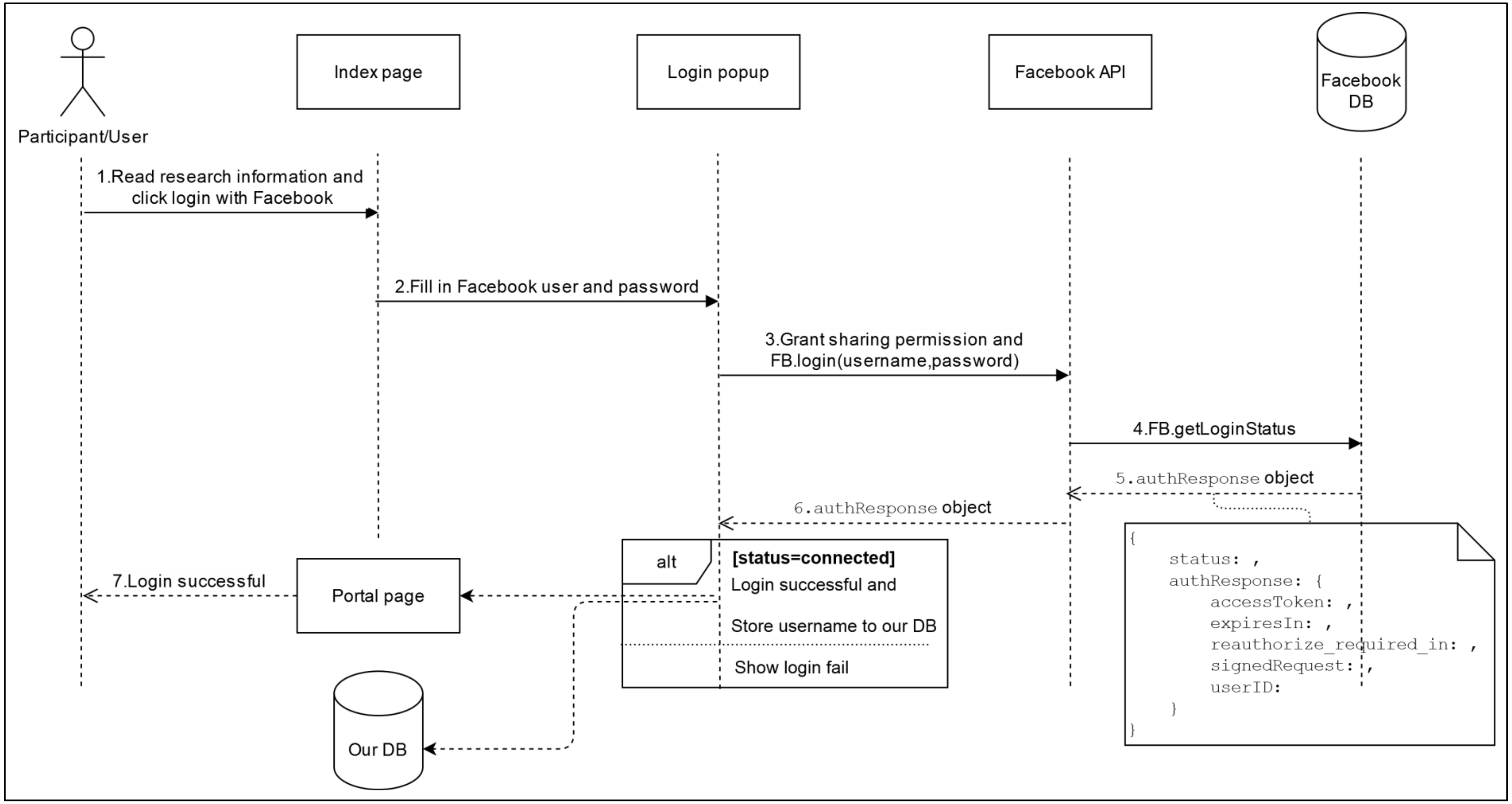
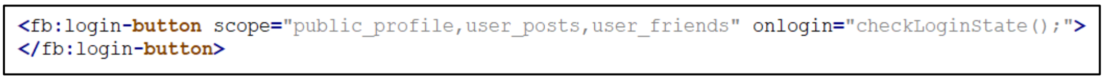


*Figure B-7 The sequence diagram of the technical details of collecting data from a social network platform.*

The "Login with Facebook" button is embedded with the code shown Figure B-8. This asks a user for permissions for the use of public profile, friend list, and user post. The "checkLoginState" will be called when a user clicks the button.



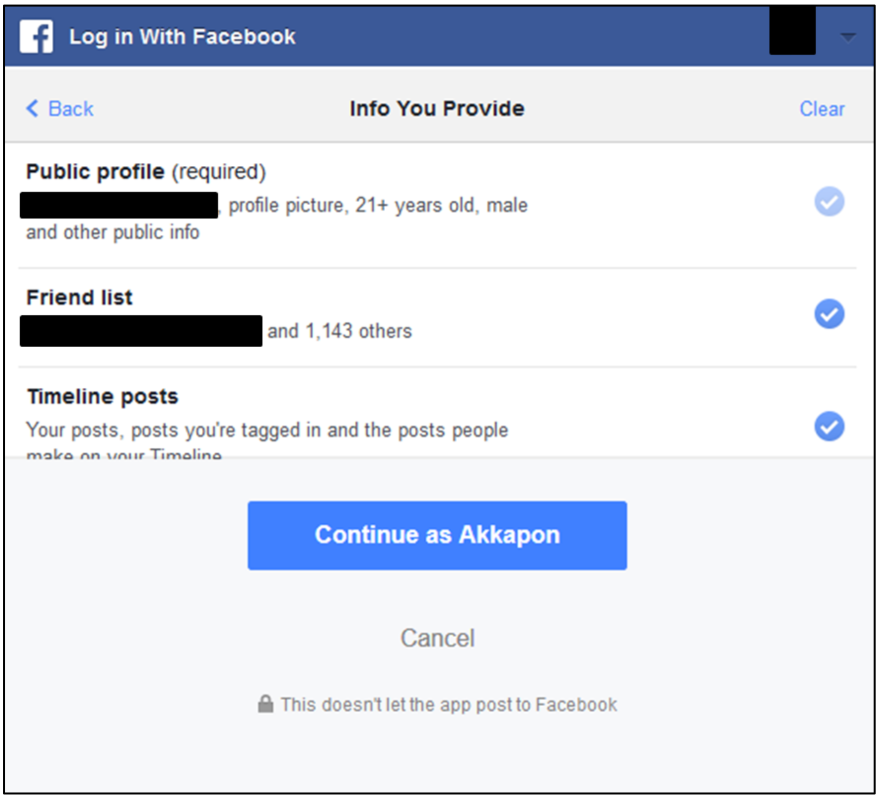
```
<fb:login-button scope="public_profile,user_posts,user_friends" onlogin="checkLoginState();">
</fb:login-button>
```

*Figure B-8 The code of login button (taken from Facebook for Developers[20]).*

The user agreeing with the consent clicks "Login with Facebook" button. The app will prompt the user to provide permissions to access their data specified in the button. The user can select which permissions are preferred to share. Figure B-9 manifests the screenshot of a user granting permissions for the use of his Facebook data. After the user click the button, the application will check a login status via FB.getLoginStatus (a Facebook login API) (see Figure B-10).

With a successful login, Facebook will return a "connected" status, redirecting the user to our portal page of our application and storing the details the user in our database.

[20] https://developers.facebook.com/docs/facebook-login/web



**Public profile** (required)

[redacted], profile picture, 21+ years old, male and other public info ✓

**Friend list**

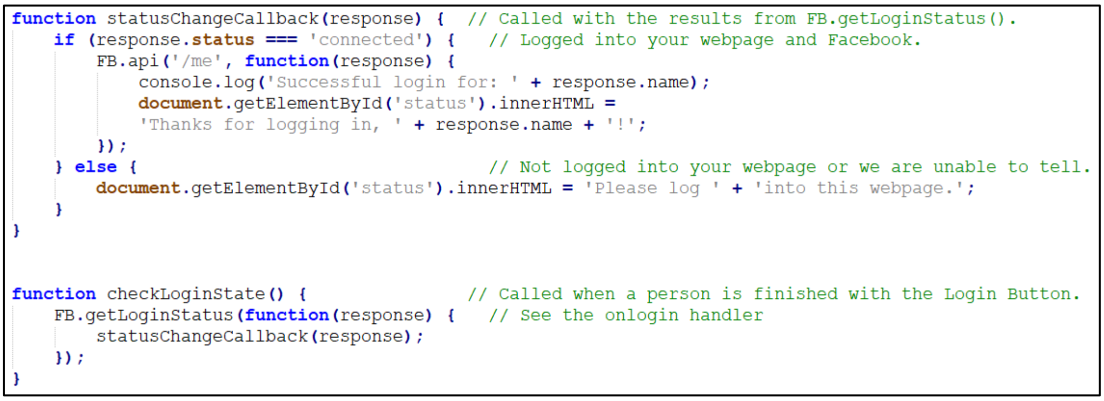
[redacted] and 1,143 others ✓

**Timeline posts**

Your posts, posts you're tagged in and the posts people make on your Timeline ✓

This doesn't let the app post to Facebook

*Figure B-9 The screenshot of a pop up to ask for permissions to access data from participants.*



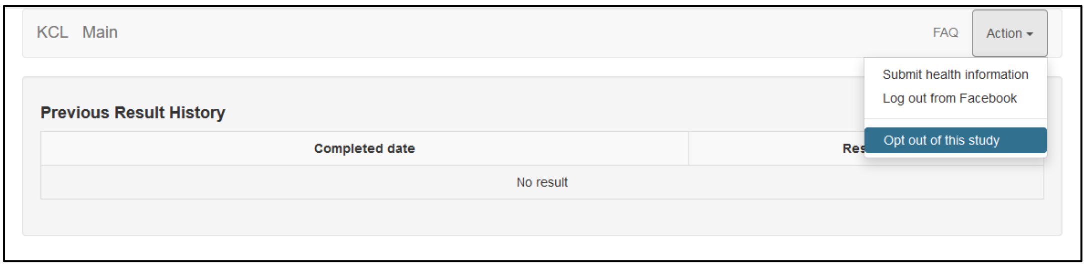
```
function statusChangeCallback(response) {  // Called with the results from FB.getLoginStatus().
    if (response.status === 'connected') {   // Logged into your webpage and Facebook.
        FB.api('/me', function(response) {
            console.log('Successful login for: ' + response.name);
            document.getElementById('status').innerHTML =
            'Thanks for logging in, ' + response.name + '!';
        });
    } else {                                 // Not logged into your webpage or we are unable to tell.
        document.getElementById('status').innerHTML = 'Please log ' + 'into this webpage.';
    }
}


function checkLoginState() {               // Called when a person is finished with the Login Button.
    FB.getLoginStatus(function(response) {   // See the onlogin handler
        statusChangeCallback(response);
    });
}
```

*Figure B-10 The code for logging in a user profile with granted permisssion (taken from Facebook for Developers[21]).*

[21] https://developers.facebook.com/docs/facebook-login/web

### B.2.1.2 User Portal

Successfully logging in, the application will redirect a user to a portal page (Figure B-11) in which the user can submit new health information, see a previous health result, and opt out from our study.



Previous Result History

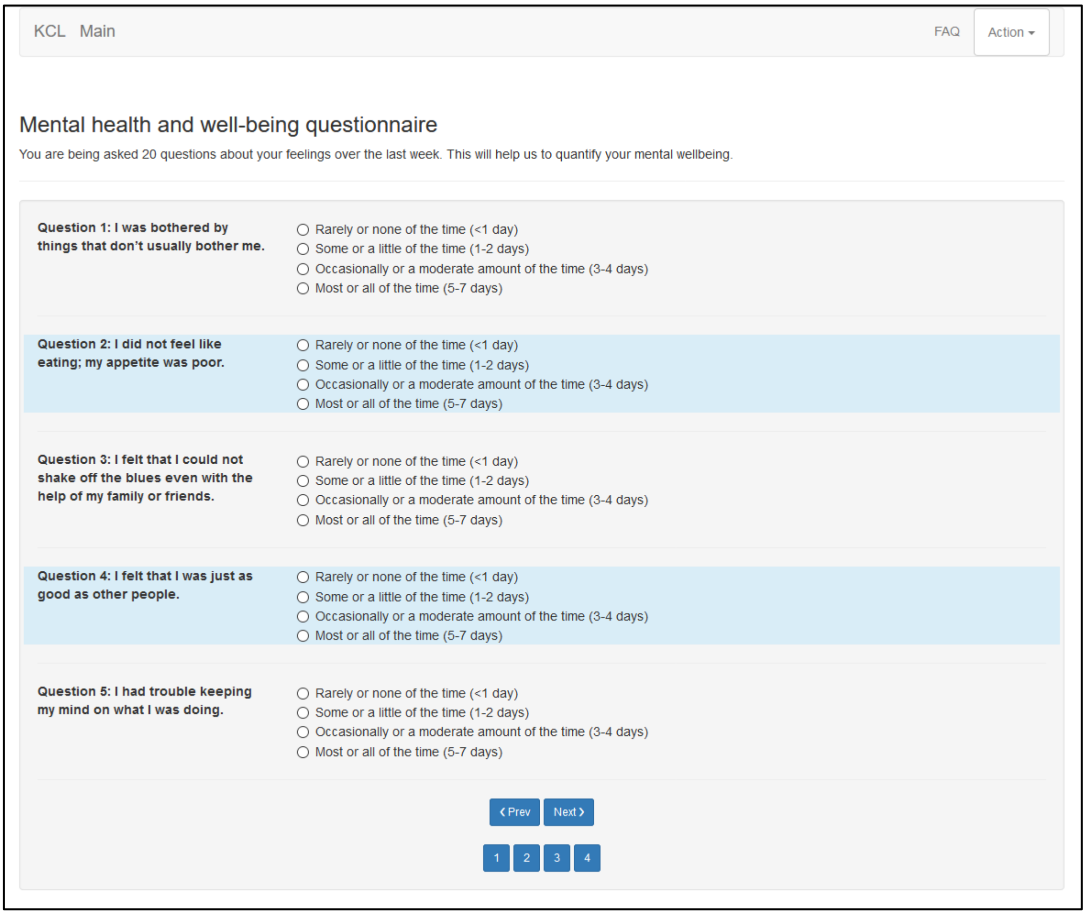
| Completed date | Res |
|---|---|
| No result | |

*Figure B-11 The screenshot of the user portal.*

We provide an opt-out option for a user to be compliant with Facebook regulations[22] and GDPR guidelines. When a user opts out from our study, all data of the user will be removed from our study.

The application will open CES-D to screen their symptoms of depression when the user clicks the “Submit health information” button. Figure B-12 shows the health screening page including 20 questions. The answers from the questions will be stored in our database after completing the questionnaire. The user can explore their previous screening results. Figure B-13 illustrates the screenshot of the main portal page with a screening result.

[22] https://developers.facebook.com/docs/facebook-login/overview



Mental health and well-being questionnaire

You are being asked 20 questions about your feelings over the last week. This will help us to quantify your mental wellbeing.

**Question 1: I was bothered by things that don't usually bother me.**

○ Rarely or none of the time (<1 day)
○ Some or a little of the time (1-2 days)
○ Occasionally or a moderate amount of the time (3-4 days)
○ Most or all of the time (5-7 days)

**Question 2: I did not feel like eating; my appetite was poor.**

○ Rarely or none of the time (<1 day)
○ Some or a little of the time (1-2 days)
○ Occasionally or a moderate amount of the time (3-4 days)
○ Most or all of the time (5-7 days)

**Question 3: I felt that I could not shake off the blues even with the help of my family or friends.**

○ Rarely or none of the time (<1 day)
○ Some or a little of the time (1-2 days)
○ Occasionally or a moderate amount of the time (3-4 days)
○ Most or all of the time (5-7 days)

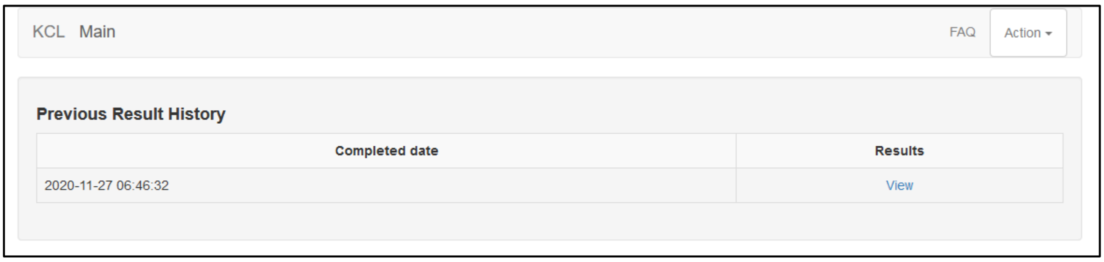
**Question 4: I felt that I was just as good as other people.**

○ Rarely or none of the time (<1 day)
○ Some or a little of the time (1-2 days)
○ Occasionally or a moderate amount of the time (3-4 days)
○ Most or all of the time (5-7 days)

**Question 5: I had trouble keeping my mind on what I was doing.**

○ Rarely or none of the time (<1 day)
○ Some or a little of the time (1-2 days)
○ Occasionally or a moderate amount of the time (3-4 days)
○ Most or all of the time (5-7 days)

*Figure B-12 The screenshot of a mental health screening page.*



**Previous Result History**

| Completed date | Results |
|---|---|
| 2020-11-27 06:46:32 | View |

*Figure B-13 The screenshot of the user portal page with a previous screening result.*

## B.2.2 Data Capture tool for Research Members

In the previous section, we introduced the data gathering application for participant use. This will explain the data capture application for research members. Figure B-14 displays the process of technical details of the data capture application. Research members can observe the number of participants and load data from participants granting their permissions.

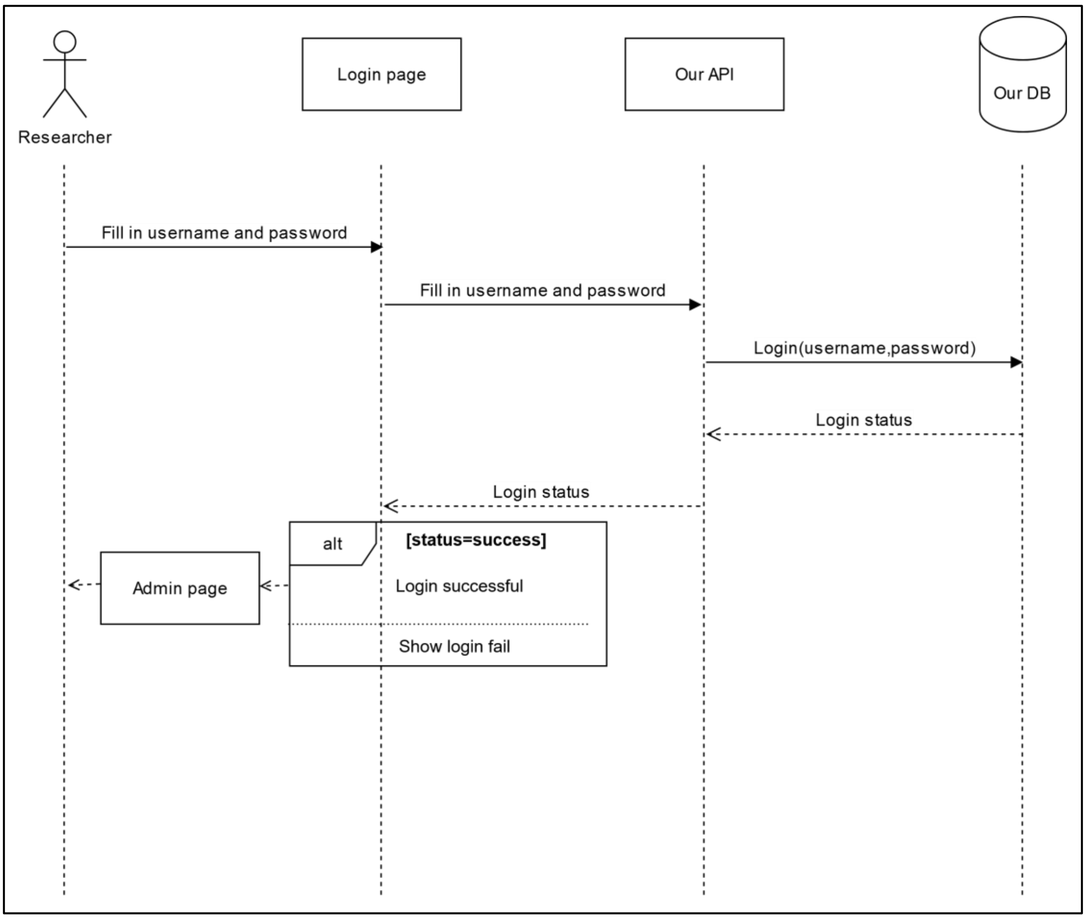
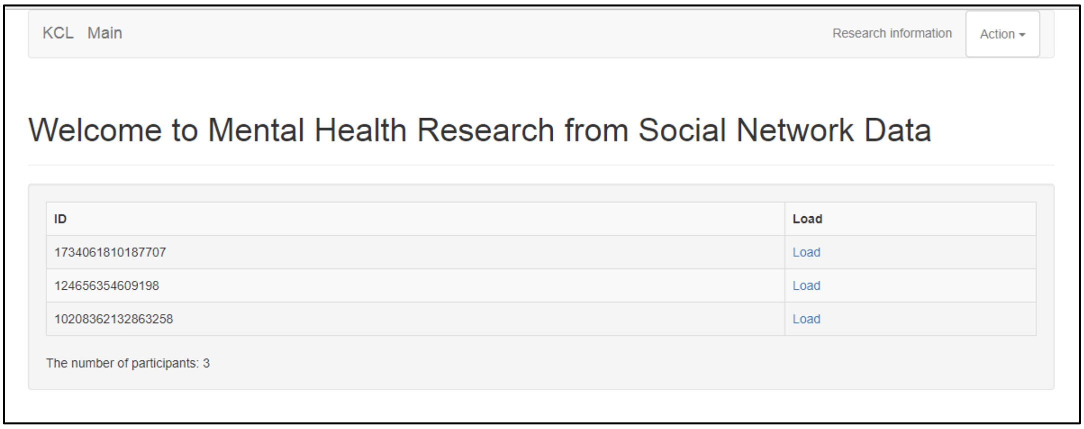


*Figure B-14 The sequence diagram of the technical details of research member application.*

Prior to the use of the data capture application for researchers, a member needs to log in to an administrator page with his provided username and password. This login mechanism is to protect data of participants accessed by other third parties.

Figure B-15 presents the administrator page with the list of test users. When a research clicks the “Load” button, the application will connect to the “User Posts” API to retrieve posts on a specific user ID. Figure B-16 shows the example code of capturing user posts. The API will return user posts as JSON format shown in Figure B-17.

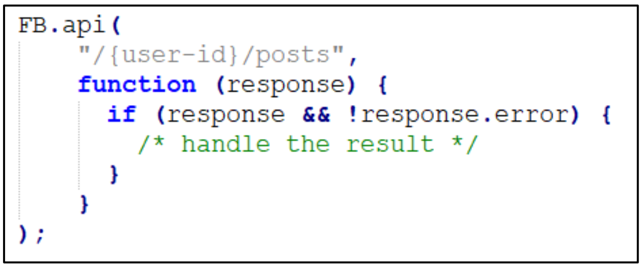


Welcome to Mental Health Research from Social Network Data

| ID | Load |
|---|---|
| 1734061810187707 | Load |
| 124656354609198 | Load |
| 10208362132863258 | Load |

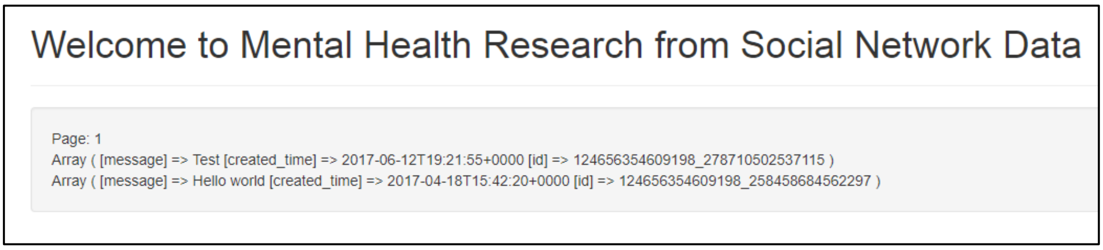
The number of participants: 3

*Figure B-15 The screenshot of the administrator page to observe the number of participants and obtain data from participants granting their permissions.*



```
FB.api(
    "/{user-id}/posts",
    function (response) {
      if (response && !response.error) {
        /* handle the result */
      }
    }
);
```

*Figure B-16 The code of retrieving user posts (taken from Facebook for Developers[23]).*



Welcome to Mental Health Research from Social Network Data

```
Page: 1
Array ( [message] => Test [created_time] => 2017-06-12T19:21:55+0000 [id] => 124656354609198_278710502537115 )
Array ( [message] => Hello world [created_time] => 2017-04-18T15:42:20+0000 [id] => 124656354609198_258458684562297 )
```

*Figure B-17 The example of retrieving posts from a test user.*

## B.2.3 Database Schema

After providing the technical details of a data collection application for participants and research member, this will describe the details of our database schema. Figure B-18 presents our schema including user, post, CES-D, and researcher tables. The user table stores the information of Facebook IDs of participants and participation date. The responses to the questionnaire from participants will be saved in the CES-D table. The posts of participants will be kept in the post table, when a research member clicks the "Load" button. Researcher information is stored in the researcher table.

[23] https://developers.facebook.com/docs/graph-api/reference/v9.0/user/posts

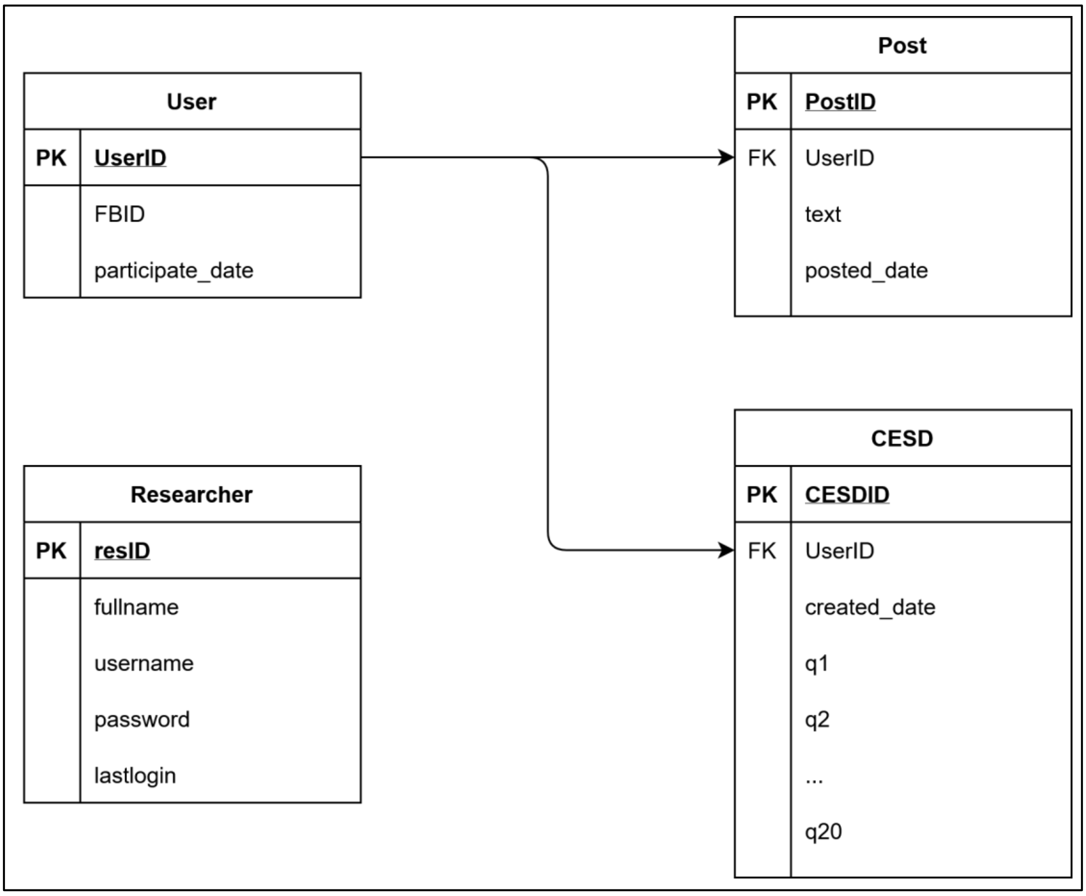


*Figure B-18 The database schema of our data capture application.*

## B.3 Summary

This chapter provided the technical details of our data collection toolkit from a microblogging and a social network platform. To develop a data gathering tool, we need to be compliant with regulations of the social media platform. In terms of technical practices, we need to work with secure connections between the application and social media platforms APIs to protect the participant’s data from unauthorised access by third parties.