Predicting Depression and Suicide Ideation in the Canadian Population Using Social Media Data

by

Ruba Skaik

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for the
Ph.D. degree in Computer Science

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

Abstract

The economic burden of mental illness costs Canada billions of dollars every year. Millions of people suffer from mental illness, and only a fraction receives adequate treatment. Identifying people with mental illness requires initiation from those in need, available medical services, and professional experts' time allocation. These resources might not be available all the time. The common practice is to rely on clinical data, which is generally collected after the illness is developed and reported. Moreover, such clinical data is incomplete and hard to obtain. An alternative data source is conducting surveys through phone calls, interviews, or mail, but this is costly and time-consuming. Social media analysis has brought advances in leveraging population data to understand mental health problems. Thus, analyzing social media posts can be an essential alternative for identifying mental disorders throughout the Canadian population. Big data research of social media may also endorse standard surveillance approaches and provide decision-makers with usable information. More precisely, social media analysis has shown promising results for public health assessment and monitoring. In this research, we explore the task of automatically analysing social media textual data using Natural Language Processing (NLP) and Machine Learning (ML) techniques to detect signs of mental health disorders that need attention, such as depression and suicide ideation. Considering the lack of comprehensive annotated data in this field, we propose a methodology for transfer learning to utilize the information hidden in a training sample and leverage it on a different dataset to choose the best-generalized model to be applied at the population level. We also present evidence that ML models designed to predict suicide ideation using Reddit data can utilize the knowledge they encoded to make predictions on Twitter data, even though the two platforms differ in the purpose, structure, and limitations. In our proposed models, we use feature engineering with supervised machine learning algorithms (such as SVM, LR, RF, XGBoost, and GBDT), and we compare their results with those of deep learning algorithms (such as LSTM, Bi-LSTM, and CNNs). We adopt the CNN model for depression classification that obtained the highest F1-score on the test dataset (0.898) and 0.941 recall. This model is later used to estimate the depression level of the population. For suicide ideation detection, we used the CNN model with pre-trained fastText word embeddings and linguistic features (LIWC). The model achieved an F1-score of 0.936 and a recall of 0.88 to predict suicide ideation at the user-level on the test set.

To compare our models' predictions with official statics, we used 2015-2016 population-based Canadian Community Health Survey (CCHS) on Mental Health and Well-being conducted by Statistics Canada. The data is used to estimate depression and suicidality in Canadian provinces and territories.

For depression, (n=53,050) respondents filled in the Patient Health Questionnaire-9 (PHQ-9) from 8 provinces/territories. Each survey respondent with a score $\geq$ 10 on the PHQ-9 was interpreted as having moderate to severe depression because this score is frequently used as a screening cut-point. The weighted percentage of depression prevalence during 2015 for females and males of the age between 15 to 75 was 11.5% and 8.1%, respectively (with 54.2% females and 45.8% males). Our model was applied on a population-representative dataset that contains 24,251 Twitter users who posted 1,735,200 tweets during 2015 with a Pearson correlation of 0.88 for both sex and age within the seven provinces and NT territory included in the CCHS. An age correlation of 0.95 was calculated for age and sex (separately) and our model estimated that 10% of the sample dataset has evidence of depression (58.3% females and 41.7% males).

For the second task, suicide ideation, Statistics Canada (2015) estimated the total number of people who reported serious suicidal thoughts as 3,396,700 persons, i.e., 9.514% of the total population, whereas our models estimated 10.6% of the population sample were at risk of suicide ideation (59% females and 41% males). The Pearson correlation coefficients between the actual suicide ideation within the last 12 months and the predicted model for each province per age, sex, and both more than 0.62, which indicates a reasonable correlation.

Acknowledgements

First of all, I am blessed to have Professor Diana Inkpen as my supervisor. I want to thank her for being an amazing advisor. I benefited not only from her profound insights and knowledge but also from her patience, kindness, and persistence. I am grateful for all the support provided by her, including Tamale seminars and group meetings. It is an honor working with and learning from her.

I want to express my gratitude to the Ph.D. committee members, Professor Marina Sokolova, Professor Olga Baysal, and Professor Hussein Al-Osman, for their insightful comments and feedback on my thesis.

Special thanks to Dr. Kenton White, Zunaira Jamil, Shainen Davidson, and Peter Glossop for providing me with the Twitter data for population analysis through Advanced Symboic Inc. Thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC) for funding my research. Thanks to Prasadith Buddhitha, Haifa Al-Harithi, and all TAMALE group members.

It was a long journey, and I am indebted to many people throughout these years. I want to thank my family, especially my mother, for believing in me and fighting for my dream. Thanks to my beloved husband, Adel; without you this journey would neither have started nor have ended. Thanks to my kids Ahmad and Noor for their understanding. Thanks to my sisters and brothers, for their encouragement. Thanks to my uncle for his love. Special thanks to Lama Skaik, Mohammed Sadeq, and Huda Al-Atari for their constant support and care.

Finally, thanks to all my friends, especially Reem Al-Halimi, Mariam Al-Otaibi, Maha Abu-Sharkh, Anas Nayfeh, Asmaa Abu-Bakr, and Manal Al-Cheikh.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Mental health is an area of health with one of the most substantial gaps between the seriousness of the problem and the lack information we have (Paul and Dredze, 2017). Mental illness is a leading cause of disability worldwide as reported by WHO (2019). Mental illness includes mood or personality disorders such as depression, insomnia, bipolar, schizophrenia, anxiety, and substance dependency. Millions of people suffer from mental illness and only a fraction receives adequate treatment (Puyat et al., 2016). It is estimated that 792 million people worldwide had a mental disorder in 2017 (Ritchie and Roser, 2018). As per Statistics Canada, 12.6% adults experience a mood disorder (mostly depression) during their lifetime (Public Health Agency of Canada, 2015). Furthermore, the resulting financial cost associated with mental illness in Canada is estimated at $51 billion per year (Canada, 2017). This includes health care costs and reductions in quality of life.

Depression is one of the most widely recognized mental disorders in the world. It is the main contributor to non-fatal prevalence of impairment, accountable for a high number of disability-adjusted years of life globally (Public Health Agency of Canada, 2015) and one of the leading causes for suicide. Early recognition of signs of depression and application of the proper treatments can make a difference for 80% of affected people[1] allowing them to cope and get back to their daily routines and regular activities.

[1] https://cmha.ca/fast-facts-about-mental-illness

Although "suicide is not a mental illness" according to the Canadian Mental Health Association (CMHA) 2017, but it is considered as a serious public health problem and the right prompt response can mitigate it. Most suicide attempts are ascribed to mental disorders, very commonly depression. Suicide has become one of the leading causes of death worldwide and is increasingly threatening the lives of children. For instance, the rate of youth suicide in Canada is the third highest in the developed world[2]. Every year about 800,000 people die from suicide. According to Statistics Canada, 4,012 people decided to take their own lives in Canada in 2019, at a rate of 9 per 100,000 people. Each suicide case has significant implications on the physical and emotional well-being of families and societies in general. An early detection of suicide ideation can prevent many cases of suicide and help in identifying those that need immediate counseling. This is one of the crucial steps of maintaining global mental health.

Figure 1.1 shows the numbers estimated by the Institute for Health Metrics and Evaluation and reported in (GBD, 2017) of conditions of mental health and drug use world wide between males and females[3]. While mental health and substance use problems are significantly under-reported, this is true across all countries, but specifically at lower income levels where data is more difficult to obtain, and mental health disorders receive less intervention. It is not unexpected that the amount of the research done to identify mental health disorders is proportional to the importance of the mental health disorders.

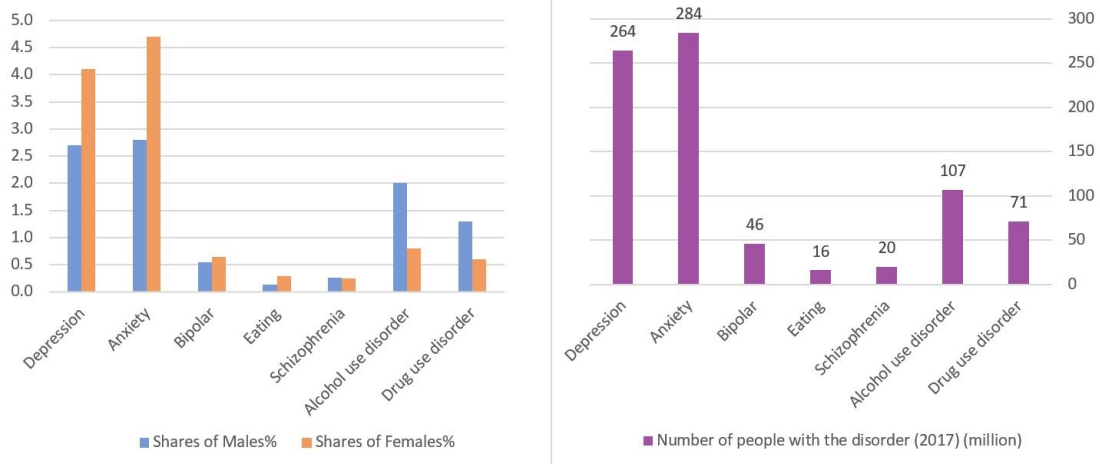

Figure 1.1. The share of global population with disorders for 2017

[2]https://cmha.ca/fast-facts-about-mental-illness

[3]https://ourworldindata.org/mental-health#prevalence-of-mental-health-and-substance-use-disorders

1.2 Motivation

Public health surveillance is "the ongoing, systematic collection, analysis, interpretation, and dissemination of data regarding a health-related event for use in public health action to reduce morbidity and mortality and to improve health" (German et al., 2001). A critical requirement for a surveillance system is to obtain reliable information and evidence in a timely manner. Identifying early warning signs in patients can lead to prompt medical treatments to avoid relapse and hospitalization (Abdullah and Choudhury, 2018). Furthermore, identifying suffering clusters in terms of demographic information can guide governments to target each group with suitable vigilance programs, plan required medical assistance to the concerned parties in early stages, and allocate the necessary resources to reduce the encumbrance of mental illness in their respective region. Recognizing people with mental illness is a challenging task. People with mental illness may not report their condition (Marcus et al., 2012). Additionally, the medical service may not be available once needed. For example, for adult mood disorders, the average wait for outpatient services in Ontario is 57 days, and inpatient services are 47 days, well beyond the suggested 28 days standard (Loebach and Ayoubzadeh, 2017).

Social media use has risen sharply over the last ten years. In 2020, there were approximately 25.35 million social media users in Canada, i.e., more than 65% of the population are social media users[4]. Social media offers advantages over traditional data sources, including ease of access, reduced cost, and up-to-date data availability. We can learn about public health topics by passively analyzing existing social media data (Paul et al., 2016). Thus, analyzing social media posts can play an important alternative in monitoring trends instantly and identifying mental disorders throughout the population. The importance of this field emerges with the progressive rise in the number of depressed users due to the COVID-19 pandemic and an anticipated increase in the number of suicides (Sher, 2020), this field becomes even more promising in providing decision-makers, such as Public Health Agency of Canada (PHAC, with rapid tools to mitigate risks.

[4]https://www.statista.com/statistics/260710/number-of-social-network-users-in-canada

1.3 Problem Statement

Our focus in this thesis is to build a generalizable classification model using the techniques of Natural Language Processing (NLP) and Machine Learning (ML) that can identify depressed and suicidal users from social media data that is representative of the Canadian population. Different methods are used to select the best features that lead the model to scale up. Using conventional machine learning algorithms and state-of-the-art deep learning architectures, we address both tasks as binary classification problems. We explore the three following major problems in this research: building a generalizable model for predicting depression using limited labeled data as explain in Section 1.3.1, using transfer learning to train a suicide ideation detection model as explained in Section 1.3.2, and finally using these models for predicting depression and suicide ideation at the population level.

1.3.1 Generalizability

In this thesis, we aim to improve the quality of depression and suicide ideation classifiers at user-level, relying on textual features within the social media users' posts. In addition, deep learning algorithms will be applied to reach this objective. The predictive models are trained using one dataset and then tested on an unseen dataset with originally different source and annotation process. The main focus is on depression and suicide ideation detection.

A major depressive disorder (MDD) is a medical condition that causes impairment of the patient's life in the social, work, or other aspects. It tends to affect mood and behaviour along with different vital functions like appetite and sleep. MDDs have five depressive symptoms which last for at least 2 consecutive weeks, according to DSM-5 and ICD-10-CM codes (International Classification of Diseases, Tenth Revision, Clinical Modification). DSM-5 is the Mental Disorders Diagnostic and Statistical Manual used by clinicians and psychiatrists. It was published 2013 by the American Psychiatric Association and updated on 2020 with new codes (ICD-10-CM) that define all categories of adult and child mental health disorders, to improve diagnosis, treatment, and research. We will experiment with different features, including word embeddings, to distinguish between depressed and non-depressed social media users and probe textual features to identify people at suicide risk.

1.3.2 Transfer Learning

Manually labeling social media users is a challenging task. It is expensive and requires dedicated time from domain experts. However, it is an essential part for training an automatic text classifier. In this research, we examine training a suicide ideation classifier using an annotated Reddit dataset and testing it on a different data set composed of tweets. Utilizing conventional machine learning and deep learning algorithms, we establish that this is feasible.

1.3.3 Population Perspective

Depression in Canada has risen to national attention, especially with the spike increase of depressed people due to COVID-19 pandemic. The precautions and restrictions to curb the spread of COVID-19 had increased the rates of suicide as well (John et al., 2020; Reger et al., 2020). Thus, these two correlated issues need urgent consideration.

The aim of this research is to predict depression and suicide ideation at the population level. For this, we start with user level classifiers, and for each social media user, we utilize an estimate with a confidence level for the location, age and sex. Then, we measure the correlation between our predictions and the data available from Statistics Canada.

1.4 Hypothesis

We hypothesise that social media text can complement conventional methods of population-driven analytical approaches for depression and suicide ideation detection. Given two different labeled sets $\mathcal{T}_1 = \{(x_1, y_1), ...(x_n, y_n)\}$ and $\mathcal{T}_2 = \{(x_1, y_1), ...(x_n, y_m)\}$ where x_i is the set of tweets or posts posted by $user_i$ and y_i is the label assigned to the user, either by an expert annotation or by self-disclosure of a mental illness. We hypothesise that based on these two sets, we can train a generalizable model capable of giving insights into the population status by applying this model to a population-representative sample of social media users and compare its results with the official statistics.

1.5 Contributions

Below are the key contributions of this thesis:

- A pioneering study using social media text for depression and suicide ideation population estimation with reference to national surveys at provincial and territorial level, sex, and age.
- Comparing traditional and deep learning algorithms generalizability at the user level then at the population-level.
- Suicide ideation prediction at the user level utilizing the information gained from depression models features engineering then adding discriminating engineered features to the word embeddings in a deep neural network (CNN) for better "at-risk" identification.
- Depression and suicide ideation prediction on the Canadian population-level and comparing the models' prediction with the Canadian Community Health Survey (CCHS) on Mental Health and Well-being conducted by Statistics Canada.

1.6 Thesis Organization

The remainder of this thesis is organized as follows.

- In Chapter 2, we provide the essential background including the levels of analysis and related work.
- Chapter 3 explains the data collection and annotation methods in the literature, followed by a description of the depression, suicide ideation and population datasets that are used in this research.
- Chapter 4 summarizes the methods for classifying Twitter users into depressed or not depressed users using traditional and deep learning ML methods and applies the best model - in terms of the recall anf F1-score metric - for population level predictions, then compares the results with the national cross-sectional Canadian

Community Health Survey (CCHS) that provides health information at the regional and provincial levels.

- Chapter 5 describes the methods for detecting suicide ideation and evaluates our proposed approaches' results; then, we apply the best-performing model to predict suicidal thoughts at the population level. The prediction results are then compared with the (CCHS) across Canada's provinces and territories.

- Chapter 6 concludes the thesis, discusses limitations and challenges, and presents directions of future work.

1.7 Published Papers

(1) Ruba Skaik and Diana Inkpen. "Using Social Media for Mental Health Surveillance: A Review". ACM Computing Surveys, 53(6), December 2020.

(2) Ruba Skaik and Diana Inkpen. "Using Twitter Social Media for Depression Detection in the Canadian Population". In Proceedings of the 3rd Artificial Intelligence and Cloud Computing Conference (AICCC 2020), AICCC 2020, pages 109—-114, Kyoto, Japan, 2020.

(3) Ruba Skaik and Diana Inkpen. "Suicide Ideation Estimators within Canadian Provinces using Machine Learning Tools on Social Media Text". Journal of Advances in Information Technology. Selected papers from the 13th International Conference on Machine Learning and Computing (ICMLC), 2020.

(4) Diana Inkpen, Ruba Skaik, Prasadith Kirinde Gamaarachchige, Dimo Angelov and Maxwell Thomas Fredenburgh. "uOttawa at eRisk 2021: Automatic Filling of the Beck's Depression Inventory Questionnaire using Deep Learning".

Chapter 2

Related Work

2.1 Overview

This chapter[1] explains the relevant literature for depression and suicide ideation detection on post-level, user-level and population-level preceded with an introduction to the features that are commonly used in this field.

2.2 Features for Social Media Text

Social media data can provide a valuable source of information for public health research. Understanding data and the domain of discourse is vital for building a good model. Accordingly, detecting mental disorders using social media posts requires a thorough understanding of the key predictors of the illness, called features in ML terminology. Many researchers tried to determine the contributing features utilizing different NLP approaches to build an accurate predictive model.

Most predictive models focus on determining the best features that contribute to the problem under analysis in order to design good classifiers. Selecting the best set of features that help to reduce the dimension from the dataset is influential to the learning process. Because of the heterogeneity of social media content, a variety of features can be developed

[1]Parts of this chapter are published in (Skaik and Inkpen, 2020b)

starting from textual and linguistic, to user-based and metadata related features (Wijeratne et al., 2017). As mentioned earlier, only a subset of these features has the required ability to distinguish classes for specific applications and contexts, in particular, to predict user types and behaviors (Kursuncu et al., 2018). Feature engineering or extraction aims to reduce the number of features extracted from the dataset under study by choosing the most discriminative features. Reducing data dimensionality via the feature extraction process helps avoid the curse of dimensionality (Cummins et al., 2015). The main target is to identify strongly relevant and non-redundant features (Li et al., 2017) which is a challenging task. Using deep learning frameworks help capture related features during the learning process without exhaustive feature engineering (Orabi et al., 2018). Table 2.1 shows the summary of the features that have been used in various mental health signs analysis studies. Several efforts have been attempted to predict mental illness within social media content on post-level, user-level, or population-level. In this section, more details and references on the analysis levels will be explained.

2.3 Post-level Analysis

Predicting indicators of mental disorders within posts could be an intermediate step towards a more comprehensive model (Lin et al., 2017; Wang et al., 2017b). As discussed in more depth in the following sections, some studies started with post-level to reach higher-level analysis (Tadesse et al., 2019; Gaur et al., 2018; Tsugawa et al., 2015). Others focused on posts, either to binary classify the post for its inclusion of symptoms of mental illness or to quantify the signs of mental illness in terms of significance and urgency to triage the post to the appropriate level of importance, such as (Gui et al., 2019; Howard et al., 2020).

Figure 2.1 shows the difference between depression and non-depression contents in post level. The post itself has either explicit or implicit attributes. The explicit attributes are the raw attributes provided from the social media framework such as the post itself or the metadata embedded within the post, like the time of the post, the number of up-votes/shares/replies (based on the application) and location of the post for enabled GPS-tagging. On the other hand, the implicit attributes can be inferred from explicit attribute(s) with a simple or more complicated process, such as the post sentiment, emotions,

Table 2.1. List of features used for mental health classification using ML

Type	Description	References
Lexical	Short/long words ratio, word frequency, LIWC lexicon, Bag of Words, n-grams, words that show surprise, exaggeration or emphasis.	(Muderrisoglu et al., 2018; Burnap et al., 2017; Chen et al., 2018b; De Choudhury et al., 2017; Colombo et al., 2016; Coppersmith et al., 2015c; Doan et al., 2017; Huang et al., 2014; Karmen et al., 2015; Kavuluru et al., 2016; Kumar et al., 2015; Mowery et al., 2017a; Nguyen et al., 2014, 2017a; Reece et al., 2017; Yang and Mu, 2015)
Syntactic	POS, verb tenses, first-person pronouns, usage of intensifier terms, dependency relation, emotion and sentiment analysis.	(Burnap et al., 2015; Kang et al., 2016; Kumar et al., 2015; Mowery et al., 2017b; Wang et al., 2018; Acuña Caicedo et al., 2020)
Social networking	Type & number of connections with other users as inlinks or outlinks.	(Colombo et al., 2016; De Choudhury et al., 2013b; Johnson Vioulès et al., 2018; Kumar et al., 2015; Park et al., 2013)
Pattern of Life	The behavior of the user during a specific period concerning the volume and time of posts and the type and number of connections with other users.	(Chen et al., 2018a; Coppersmith et al., 2014a; Nguyen et al., 2017b; Peng et al., 2017)
Demographics	The users' demographic data, such as age, gender, ethnicity, income, education, and personality.	(Chen et al., 2015; Nsoesie et al., 2016; Oh et al., 2017; Peng et al., 2017; Preot et al., 2015; Schwartz et al., 2013b)
Word embedding	Representing users' posts vocabulary and capturing the semantic and syntactic relations with other words.	(Amir et al., 2017; Joshi et al., 2018; Kim et al., 2016; Lin et al., 2016; Orabi et al., 2018)
Topic modeling	Identifying topic patterns that are present across the users' posts.	(Chen et al., 2015; Cohan et al., 2017; Nguyen et al., 2017b; Resnik et al., 2013, 2015b; Toulis and Golab, 2017; Seah and Shim, 2018)

sleeping patterns and the type of the post based on textual analysis including the use of contextualized embeddings that encode aspects of the users' posts.

Figure 2.1. Word cloud overview of emerging terms for control people (left) and depressed people (right) based on self-reporting depression labeled dataset by Shen et al. (2017)

.

2.4 User-level Analysis

For user-level classification, multiple posts are aggregated as a single document or behavioral changes are detected over a defined period. This can be done in a hierarchical manner starting from the post-level using only the posts content (Muderrisoglu et al., 2018; Amir et al., 2017; Orabi et al., 2018; Schwartz et al., 2014; Zhang et al., 2015), or by considering user-defined or derived information such as gender and personality (Preot et al., 2015), or by adding behavior patterns and social engagements (Lin et al., 2014; Shuai et al., 2018; Yates et al., 2017).

For example, using multi-level dual-context language and BERT Matero et al. (2019) measured the suicide risk within Reddit posts. Sekulić and Strube (2019) utilized a Hierarchical Attention Network (HAN) on word and sentence level to distinguish posts and expressions that are valuable for a binary classification task to anticipate if a user experiences one of nine distinct disorders (Depression, ADHD, Anxiety, Bipolar, PTSD, Autism, OCD, Schizophrenia and Eating disorder).

Ramírez-Cifuentes et al. (2020) used lexical, behavioural, sentiment, social networking and psychological features based on Twitter textual in addition to image-based scores to distinguish users at suicide risk. They propose SPVC (short profile version classifier) which selects a subset of suicide ideation-related tweets based on the highest probability values.

2.5 Population-level Analysis

Studies show the usefulness of using social media data to recognise mental health indicators in countries as a tool of public health surveillance, but it is necessary to support the accuracy of social media results from other sources, such as national surveys. There are two main approaches for textual analysis of social media content for insights of mental health issues for population-level mental-illness detection:

- **Bottom-up Approach**: Using this approach, the researchers start with individual models, then generalize them to make an inference about the population (Coppersmith et al., 2014a, 2015c, 2014b; Tech et al., 2017; Yazdavar et al., 2017). Figure 2.2[2] illustrates the steps that need to be followed to conclude a population-related inference. Importantly, the sample under study has a significant impact on the findings because often only major classes are represented, excluding crucial minority classes. Hence, researchers had utilized techniques typically used in regular surveys. To have a population-representative sample, researchers had utilized different techniques to rectify representation errors, such as probability sampling techniques including Stratified Sampling (De Choudhury et al., 2016; Wang et al., 2019a), Simple Random Sampling (Cheng et al., 2017; Liu et al., 2017; Calderon-Vilca et al., 2017; Shing et al., 2018), Cluster Sampling (Schwartz et al., 2013a) or Multi-Stage Sampling (De Choudhury et al., 2013c). In addition, researchers had used non-probability sampling techniques such as snowball sampling to deal with minority classes (Balani and De Choudhury, 2015; Tsugawa et al., 2015; Wee et al., 2017; Wolohan et al., 2018; Zhao et al., 2018).

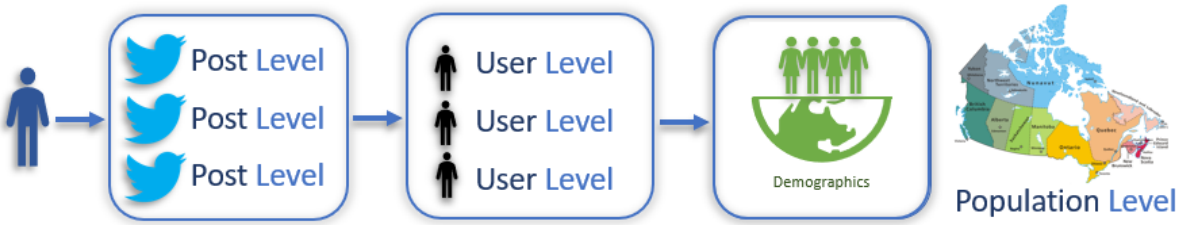

Figure 2.2. Bottom-up approach for population textual analysis

.

[2]Map: https://en.wikipedia.org/wiki/Provinces_and_territories_of_Canada

- **Top-down approach**: In this approach, studies use the aggregated data of the population, then make inference from that data (Coppersmith et al., 2017; Culotta, 2014; Schwartz et al., 2013a; Jaques et al., 2016; Gruebner et al., 2017a; Nguyen et al., 2017b). This type should be applied carefully, otherwise the differences among the subgroups could be masked and dissolved during the aggregation process shown in Figure 2.3.

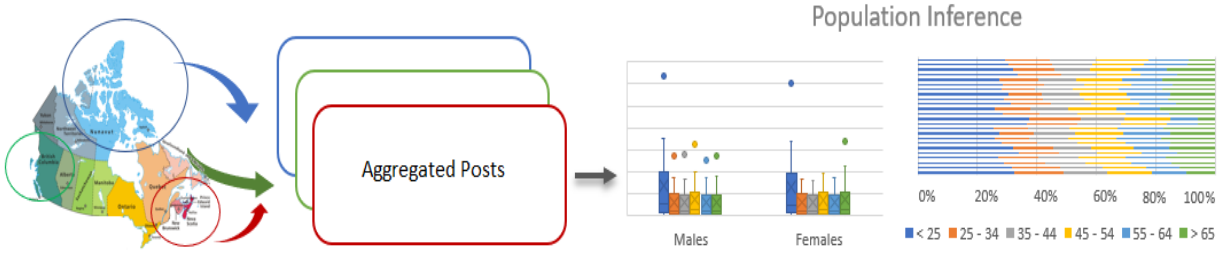

Figure 2.3. Population top-down approach using social media text

Machine learning algorithms play a vital role for modeling relationships between features (Hahn et al., 2017). It is well known that there is no universal algorithm for classification. It is increasingly important to understand the diverse cultures and societies contributing to social media. Adding demographic attributes including age, gender, and personality type affect predicting health status (Preot et al., 2015). Religion, ethnicity, marital and socioeconomic status are expected to add value in the research. Previous researches done on users' posts and meta data have shown that demographic information can be extracted by applying different ML algorithms with an accuracy ranging from 60% to 90% (Sinnenberg et al., 2017), or can be extracted from profile information, such as username, screen name, biography, and profile image with Macro-F1 measure 0.9 for gender and Macro-F1 measure 0.5 for age (Wang et al., 2019b).

Mental illness classification using machine learning utilizes selected features to categorize data instances into n classes of mental conditions, each instance can have m labels. In its simplest form; n=2 and m=1. The classifier learns from the labeled training examples as shown in Figure 2.4. Supervised learning algorithms such as SVM, Naive Bayes (NB), K-Nearest Neighbors (KNN), Logistic Regression (LR), and Decision Trees (DT) are em-

ployed. Recently, deep learning models have been extensively applied in the field of NLP and show remarkable enhancement in the prediction process. One of the key advantages of deep neural networks is its ability to learn input representation and parameters of the neural network without the need of domain-specific feature engineering. The lower layers learn simple features and cascade its learning to higher layers that yields to identify complex relationships between the input text/post(s) and the label/diagnosis.

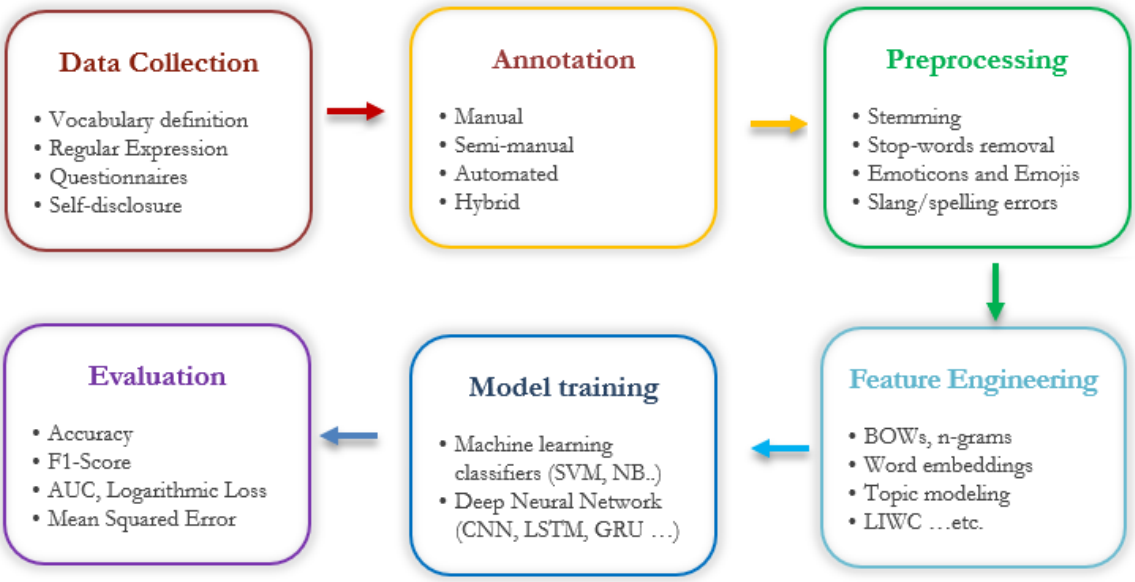

Figure 2.4. An overview of a machine learning process

Social media has been employed to assess population health statistics. For example, Culotta (2014) gathered approximately 1.4 million Twitter users spread over the 100 most crowded counties in the US and performed a linguistic analysis on tweets' text and users' description to predict 27 health-related topics ranging from lack of health insurance, obesity, teen births to an indication of mental illness. He contrasted the output of his model with figures from the County Health Rankings & Roadmaps. A significant correlation was observed with six health measures and models with added linguistically analyzed Twitter data improved predictive accuracy for 20 community health measures. In fact, the distribution of a county's textual characteristics may be insightful and predictive of the county's medical activity and health outcomes (Nguyen et al., 2017b). Moreover, a significant correlation with county-level health statistics was reported when adding word categories lin-

guistic features such as Linguistic Inquiry and Word Count (LIWC) and PERMA. LIWC[3] is a billable text analysis tool developed by psychologists across the world in their native languages that computes the usage of different categories of words in a text (Pennebaker et al., 2015). PERMA lexicon includes the vocabulary reflecting human psychological well-being factors including: Positive emotions, Engagement, Relationships, Meaning, and Accomplishment (Adler and Seligman, 2016).

Recent studies by analyzing textual features and LIWC have detected mental health conditions (Nguyen et al., 2017b,a; Johnson Vioulès et al., 2018; Kavuluru et al., 2016; Lehrman et al., 2012), personal events (Lin et al., 2016), and using topic modeling using Bayesian probabilistic modeling tools such as Latent Dirichlet Allocation (LDA) (Resnik et al., 2013; Paul and Dredze, 2014; Seah and Shim, 2018). In some cases, experiments show that reduced feature sets and simple lexical features can yield comparable results to a much larger feature dataset (Mowery et al., 2017a).

In the following subsections, we will summarize the recent techniques used in the most prominent areas of research: depression and suicide ideation in the population level.

2.5.1 Depression Detection

Due to its importance, the depression disorder has received considerable attention among researchers. NLP along with ML techniques were used on social media to detect depression (Coppersmith et al., 2014a; Karmen et al., 2015; Resnik et al., 2015b). Schwartz et al. (2014) trained a regression model on Facebook data to predict and measure the degree of depression. Also, De Choudhury et al. (2013a) developed a probabilistic model to detect the behavioral changes associated with the onset of depression. Mowery et al. (2016) used lexical and emotional features to identify depressive symptoms such as anhedonia (reduced motivation to feel pleasure), insomnia, loss of energy, etc. Nguyen et al. (2014) achieved an accuracy of 93% using a logistic regression model to classify blog posts as belonging to depression or control sets.

Shen et al. (2017) constructed a multi-modal depressive dictionary learning model (MDL) to learn the latent features of depressed and non-depressed users on Twitter from a

[3]http://liwc.wpengine.com/

joint sparse representation of emotion, visual, social network properties, user profile, topic-level, and domain-specific features and achieved 85% in F1-score. To predict user stress levels, Lin et al. (2016) used neural network-based architecture with word embeddings (WE) learned from a Weibo dataset along with stress-related keywords. Word embeddings were found to be effective features in predicting semantic similarity between different words. Table 2.2 summarizes the research findings related to predicting depression from Twitter text.

Table 2.2. Summary of the features used for predictive models for depression with the best results achieved in terms of accuracy (Acc); F1-score (F1); or recall (R)

Ref.	**Features**	**Model**	**Metric**	**Score**
De Choudhury et al. (2013a)	Lexical (swear words & 1st person pronoun)	SVM	Acc	73%
De Choudhury et al. (2013d)	Semanic and Social	SVM	Acc	70%
De Choudhury et al. (2013b)	Semantic and Social	SVM	Acc	80%
Burnap et al. (2015)	N-gram and Semantic	Ensemble	F1	0.69
Resnik et al. (2015a)	LDA, supervised Anchor, Tf-idf	SVM	Acc	86%
Tsugawa et al. (2015)	N-gram, Semantic, Social	SVM	Acc	66%
De Choudhury et al. (2016)	Interpersonal, Interaction, linguistic	LR	Acc	80%
Mowery et al. (2016)	n-grams, emotions, LIWC, age, gender	SVM	F1	0.52
Jamil et al. (2017)	Lexical, polarity, depression terms, self report	SVM	R	0.80
Peng et al. (2017)	User profile, text and behaviour	SVM	Acc	83%
Chen et al. (2018a)	LIWC, Pattern of Life, Emotions	SVM, RF	Acc	90%
Chen et al. (2018b)	Emotions, temporal, LIWC	RF	Acc	93%
Wolohan et al. (2018)	LIWC & n-gram	SVM	Acc	82%
Wongkoblap et al. (2018)	LIWC & life satisfaction	SVM	Acc	78%
Thorstad and Wolff (2019)	Tf-idf	LR	F1	0.74

2.5.2 Predicting Suicide Ideation and Self-harm

"Suicide is preventable" said Brian Mishara[4]. Suicide prevention starts with recognizing the warning signs and taking them seriously. Thus, predicting suicide ideation from social media is considered one step towards identifying affected groups based on gender, age, geographic location, or other characteristics.

In 2015, O'Dea et al. (2015) used machine learning algorithms to distinguish strongly concerning suicide-related tweets among 14,701 tweets with an accuracy of 80%. Wang et al. (2019a) performed the same on Chinese online communities.

In 2017, Benton et al. (2017b) presented a multi-task learning (MTL) model using a feed-forward neural network using character n-gram features to predict potential suicide attempt and presence of atypical mental health. Also, a framework to instantly detect suicide-related posts on Twitter was proposed by Johnson Vioulès et al. (2018) by using NLP methods that combine text-generated features based on a lexicon ensemble. Similarly, the association between suicidal ideation and the linguistic features was examined (Cheng et al., 2017; Huang et al., 2014). Also, Jashinsky et al. (2014) found a correlation between the rate of risk per tweet per state measured by the appearance of terms associated with suicide risk and state age adjusted suicide rates.

A well-known strong predictor of completed suicide is a previous suicide attempt (Rakesh, 2017) and self-harm (Robinson et al., 2015). De Choudhury et al. (2016) built a statistical approach on data from Reddit users who shifted from mental health concerns to SI. Their approach derives markers to detect this transition incident through the three cognitive psychological integrative models of suicide including thinking, ambivalence, and decision making. In 2020, using ten neural networks, Roy et al. (2020) estimated the weight of psychological factors such as stress, loneliness, burdensomeness, hopelessness, depression, anxiety, and insomnia, in addition to sentiment polarity by training a random forest model using the ten-estimated psychological metrics to predict suicide ideation within the tweets and achieved 0.88 AUC score.

Table 2.3 shows the best reported ML methods for anticipating suicide ideation on datasets that are manually annotated.

[4]Director of the Centre for Research and Intervention on Suicide and Euthanasia and a psychology professor at the Université de Québec à Montréal

Table 2.3. Summary of the features used on social media text for predictive model for suicide ideation with the best results achieved in terms of accuracy (Acc); F1-score (F1); recall (R); or Area Under the Curve (AUC)

Ref.	**Users/Posts**	**Features**	**Model**	**Metric**	**Score**
Abboute et al. (2014)	623/3,263	WEKA	NB	Acc	63%
Huang et al. (2015)	664/-	LDA	SVM	Acc	96%
O'Dea et al. (2015)	1,820/14,701	Tf-idf	SVM	Acc	80%
Braithwaite et al. (2016)	17/-	LIWC	DT	Acc	92%
Coppersmith et al. (2016)	554/-	Sentiment	LR	F1	0.53
Burnap et al. (2017)	425/-	Tf-idf	SVM	Acc	66%
Cheng et al. (2017)	976 /-	SC-LIWC	SVM	AUC	0.48
Muderrisoglu et al. (2018)	-/785	Tf-idf	SVM	F1	0.92
Coppersmith et al. (2018)	418/197,615	WE	LSTM	R	0.85
Desmet and Hoste (2018)	257/-	Gallop & BoW	SVM	F1	0.75
Roy et al. (2020)	283/512,526	NN for psychological constructs	RF	AUC	0.88

2.6 Summary

In this chapter, related work for depression detection and suicide ideation research was reviewed. Different features used in this field were illustrated. Post-level social madia text analysis was introduced followed by the social media user-level analysis. Finally, an overview of the related work on the population-level predictions was presented.

Chapter 3

Datasets

3.1 Overview

The first step to addressing mental illness is obtaining reliable information and evidence (Paul and Dredze, 2017). Having a comprehensive and accurate dataset is a critical success factor for applying ML algorithms. A gold standard is a dataset that is used to compare ML models against it (Calvo et al., 2017). Such datasets could contain only a test set on which the performance of various classifiers can be compared, but more often they include a training set as illustrated in Figure 3.1. The classifiers can be trained on the latter, though any additional labelled or unlabelled data can be used for training if desired. This chapter summarizes the data collection techniques and annotation procedures available in the field along with available gold standards followed by a detailed description of the datasets that are used in this research.

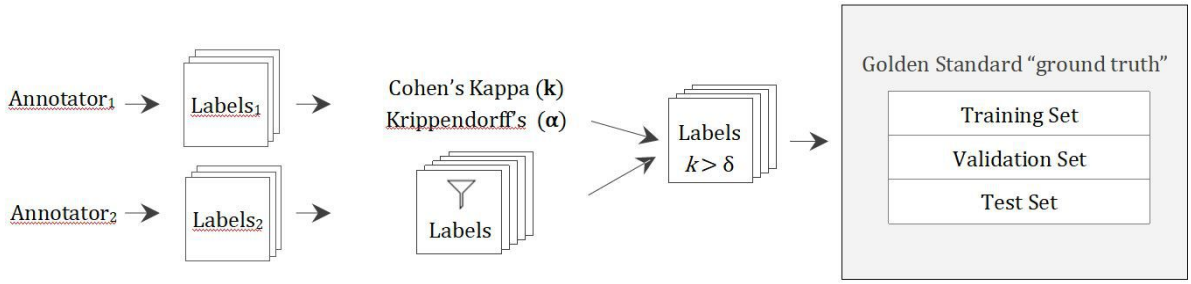

Figure 3.1. Annotation process and golden set

.

3.2 Data Collection Methods

There are several methods to gather information on social media relevant to users' mental health, including self-reporting (directly or indirectly), mental illness signs inference, manual annotations, and external statistics. Following is a description of the data collection techniques and annotation procedures:

3.2.1 Screening Surveys

Crowdsourcing platforms are considered as a significant source of explicit labels provided by human workers or volunteers. Crowdsourcing platforms like Amazon's Mechanical Turk (MTurk) or CrowdFlower enable researchers to post a questionnaire and invite people to contribute. Each participant is expected to complete a diagnostic survey with the consent of the participant to provide their social media accounts (mostly Twitter).

Different psychiatric scales are used to measure the level of the participants' mental health. For quantification of depressive symptoms, researchers may choose to use PHQ-9 (Patient Health Questionnaire) that is widely used for diagnosing and assessing the severity of depression. It measures behavioral attributes including concentration troubles, changes in sleeping habits, eating disorder, lower activity and losing interest, as well as feeling-oriented attributes such as feeling tired, down, guilt or failure as well as self-harm and suicidal thoughts. There is also the Depression Scale Center for Epidemiological Studies (CES-D) questionnaire, which provides a self-report scale of 20 multiple-choice questions designed to test depression-related symptoms such as depressed feeling, restless sleep, and decreased appetite. Similar to the latter, there are the Beck Depression Inventory (BDI) and the Short Depression-Happiness Scale (SDHS) questionnaires that serve the same purpose. For anxiety, the State Anxiety Inventory (SAI) or the Anxiety Sensitivity Index (ASI). Other questionnaires exist to measure other behaviors. For example, Life Events Checklist (LEC) is used to detect life events, Anger Rumination Scale (ARS) to measure anger. To check the presence of suicide ideation, there are many questionnaires such as Scale for Suicide Ideation (SSI), Depressive Symptom Inventory–Suicide Subscale (DSI-SS), Interpersonal Needs Questionnaire (INQ), Acquired Capability for Suicide Scale (ACSS) and many others.

Although these scales are established tools in psychiatry, sometimes it is hard to determine which scale needs to be used and is more accurate for a given population sample. For example, to predict suicide attempts, the Emotion Regulation Questionnaire (ERQ) performed better than the Scale for Suicide Ideation (SSI) questionnaire (Oh et al., 2017).

The screening method design varies depending on the study purpose and the target platform. While this method is reasonably close to clinical practice, it is expensive to administer on a large scale and suffers from sampling biases (Guntuku et al., 2017). Table 3.1 shows some examples of using such a method for data collection and its applications.

Table 3.1. Chosen studies that used screening surveys as a way to collect data about DD:Major Depression Disorder, PT: Posttraumatic Stress Disorder(PTSD), PD: Postpartum Depression and SD: Suicide

Ref.	**Field**	**Q.Type**	**Platform**	**#Users**	**#Posts**
De Choudhury et al. (2013a)	DD	CES-D	Twitter	489	69,514
De Choudhury et al. (2013d)	DD	CES-D	Twitter	476	2,157,992
De Choudhury et al. (2014)	PD	PHQ-9	Facebook	156	578,220
Tsugawa et al. (2015)	DD	CES-D	Twitter	209	574,562
Zhang et al. (2015)	SD	SPS	Weibo	697	2000/user
Braithwaite et al. (2016)	SD	DSI-SS	Twitter	135	2000/user
Reece et al. (2017)	DD	CES-D	Twitter	204	279,951
Reece et al. (2017)	PT	TSQ	Twitter	174	243,775
Almouzini et al. (2019)	DD	CES-D,PHQ-9 INQ,ACSS	Twitter	89	6,122
Stankevich et al. (2019)	DD	BDI	Vkontakte	531	32,872
Stankevich et al. (2020)	DD	BDI	Vkontakte	1,330	94,660

Some studies proposed a semi-supervised approach to mimic screening surveys and answer clinical questionnaires from social media data. Yazdavar et al. (2017) incorporated a linguistic analysis for user-generated content on social media over time for answering PHQ-9 questions. They collected a total of 23 million tweets posted by 45 thousand Twitter users with self-declared symptoms of depression in users' profile descriptions. The authors developed a probabilistic topic modeling on user tweets with partial supervision to monitor clinical depression symptoms. They achieved competitive results with a fully supervised approach with an accuracy of 68% for predicting the answers of all the questions over a time interval. Their semi-supervised topic modeling approach called (ssToT) was able to respond

with an enhanced F1 score specifically to the following symptoms: decreased satisfaction, feeling down, sleep disturbance, energy loss and appetite change. Similar approach was conducted by Karmen et al. (2015) who translated questionnaires and synonyms into a depression lexicon and used it to assign a post-level cumulative depression rating.

3.2.2 Forums Membership

Many online communities exist to discuss and support mental health topics either by the patient himself or a related person seeking help. Online communities include discussion forums, chat rooms, and blogs. Online forums are considered to be one of the essential knowledge-based resources for adults on the Internet (Korda and Itani, 2013). Support forums allow people to discuss their mental health issues and seek the required assistance from peers or specialized personnel. Some researchers rely on the users' affiliation to indicate a mental health condition. For example, Nguyen et al. (2014) considered that joining LiveJournal depression related communities is a sign for a mental illness. Furthermore, known suicide web forums such as recoveryourlife, enotalone, and endthislife are used to generate a lexicon of terms that were likely to identify suicidal thoughts within posts (Burnap et al., 2015, 2017; Colombo et al., 2016).

3.2.3 Social Media Posts

Over the last 10 years, there has been a significant increase in the usage of social media. These platforms can be utilized as a source of insights into the mental health status of the population. Some social media users share their mental illness conditions through social media platforms. Users may reveal that information as a mean to seek help from their community, challenge stigma, share their experience, and as an empowering coping mechanism (Berry et al., 2017). Social media data can be extracted using available Application Programming Interfaces (API) of the social media platforms. There are different ways of collecting and processing social media data. This is determined by the types of illnesses being researched and the platforms which are used as source material. In the following subsections, we will briefly summarize some of the most popular data sources in the social media research community.

Twitter

Twitter is the most attractive social network service among researchers. Twitter currently ranks as one of the world's leading social networks based on active users[1]. Every update the user posts to his followers on Twitter is called a tweet. Tweets are mostly accessible to the public and can be obtained and analyzed, unless flagged by the user as *"private"*. Tweets can be collected using Twitter API by searching the tweets for specific keywords, hashtags, or any defined query and can be limited to particular locations, hashtags and time periods.

Some researchers look for focal users and build the dataset incrementally based on social connections (Wang et al., 2017a; Zhao et al., 2018). Several depressive symptoms derived from mental disorders manuals such as Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) and ICD-10-CM can be identified by applying machine learning algorithms on Twitter data (Mowery et al., 2015; Prieto et al., 2014).

Other researches obtained tweets with self-reported diagnoses and filtered via regular expressions (RegEx) to capture *"I was diagnosed with ... Condition"* then the collected tweets are manually labeled by human annotators to determine whether the expression used is indicating the mentioned mental health diagnosis or not (Chen et al., 2018a; Coppersmith et al., 2014a, 2015a; Li et al., 2017; Mowery et al., 2017b; Loveys et al., 2017). For suicide related tweets, more tailored expressions need to be considered, for example `'.+(\took | \take).+\own.+\life.+'` (Burnap et al., 2017). Others extracted candidate tweets using keywords that are related to mental disorders, for example, *"Depression"* or *"Suicide"* or terms that may indicate mental disorders conditions or symptoms such as *distress, dejected, gloomy, cheerless, blue, empty, sad, feeling low, hate myself, kill myself, don't want to live anymore, ashamed of myself*, etc. (Varathan and Talib, 2014; Kale, 2015; Cavazos-Rehg et al., 2016; Mowery et al., 2017b).

There are also other ways to collect tweets that may include signs of mental disorders, like using indicative hashtags such as #MyDepressionLooksLike (Lachmar et al., 2017), #WhatYouDontSee[2], or #KMS.

[1]https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[2]https://www.buzzfeed.com/annaborges/what-you-dont-know-campaign

In addition to user generated text, Twitter includes user data and social metadata, such as geographical information and the date and time of the tweet, and user networking and interaction information. Thus, state-of-the-art applications including Twitris[3], SOCIALmetrics™ and OsoMe[4] have been built to analyze massive real-time social media data. Twitris is a linguistic social web network that utilizes user-generated social media content to understand social perspectives on real-world events. Whereas, SOCIALmetrics™ is a system that processes crawled Twitter data using NLP and text mining tools. Razak et al. (2020) presented Tweep, a rule-based system using VADER (Valence Aware Dictionary and sEntiment Reasoner) Sentiment analysis tool[5] and two machine learning algorithms: Naive Bayes (NB) and Convolutional Neural Network (CNN) to analyze tweet sentiment for logged in users and their Twitter followers. Table 3.2 presents a summary of recent research done using Twitter website as a data source.

Four of the datasets that are used in this research are Twitter-based and are explained in more details in Sections 3.4.1, 3.4.2, 3.4.3 and 3.4.5 .

Table 3.2. References for ML algorithms applied over data collected from Twitter platform on DD:Major Depression Disorder, PT: Posttraumatic Stress Disorder(PTSD), PD: Postpartum Depression, LS: Life Satisfaction, MI: Mental Illness, SI: Suicide Ideation

Ref.	**Users**	**Posts**	**Field**	**Objective**
De Choudhury et al. (2013b)	376	77,374	PD	Identify mothers at risk of postpartum depression using engagement, emotion, linguistic style and social network features.
De Choudhury et al. (2013c)	85	5(1)	PD	Predict mothers at postpartum depression risk using prenatal behavior changes in language, patterns of posting, and emotion.
Schwartz et al. (2013a)	-	82M	LS	Predict life satisfaction of US counties using linguistic features.
Abboute et al. (2014)	6,000	-	SI	Classify tweets into risky and non-risky language.

(1) months

[3]http://twitris.knoesis.org
[4]https://osome.iu.edu
[5]https://github.com/cjhutto/vaderSentiment

References for ML algorithms applied over data collected from Twitter platform by year (Cont.)

Ref.	Users	Posts	Field	Objective
Coppersmith et al. (2014a)	6,966	16.7M	MI	Analyzed language usage relevant to mental health.
Coppersmith et al. (2014b)	5,972	926K	PT	Analyzed the language usage of PTSD Twitter users utilizing LIWC.
Culotta (2014)	1.46M	4.31M	MI	Estimated health statistics for US counties using LIWC+PERMA lexicons.
Jashinsky et al. (2014)	28,088	37,717	SI	Analyzed the spatial correlation of suicide rates in the US and predicted at-risk users.
Burnap et al. (2015)	-	2000	SI	Ensemble classifier to differentiate between suicidal ideation contents and other suicide-related topics such as reporting of suicide /condolences.
Coppersmith et al. (2015c)	100(2)	-	SI	Quantifiable linguistic differences between users' posts prior to suicide attempt and control users as well as depressed users and suicide attempts in the US.
Huang et al. (2015)	7,314	-	SI	Topic modeling using extended suicide psychological lexicon.
O'Dea et al. (2015)	-	14,701	SI	Analyzed suicidality and predicted the level of concern among suicide-related tweets.
Preot et al. (2015)	1,957	6.7M	MI	Demographics & personality estimated from tweets achieved high performance to identify mental illness.
Resnik et al. (2015b,a)	2,000	3M	MI	Used sLDA to analyze linguistic signals and uncover meaningful latent structure .
Coppersmith et al. (2016)	1,088	320K	SI	An empirical study of the language trends and emotional changes for individuals before a suicide attempt.
Kang et al. (2016)	45	23,956	DD	Multi-modal analysis for mood, emoticon and images.

(2) users per state

References for ML algorithms applied over data collected from Twitter platform by year (Cont.)

Ref.	Users	Posts	Field	Objective
Mowery et al. (2016)	-	9,473	DD	Classify evidence of depression using binary features.
Benton et al. (2017b)	9,611	33.8M	MI	Multi- task learning (MTL) framework for 8 mental health conditions prediction.
Burnap et al. (2017)	-	816	SI	Differentiate between suicidal ideation contents and other suicide-related topics using Rotation Forest and a Maximum Probability voting.
Tech et al. (2017)	534,829	1.3M	MI	Explored how gender and culture influences the online conversation of mental disorder.
Jamil et al. (2017)	25,362	156,612	DD	Predict at-risk for depression users using tweet depression index.
Mowery et al. (2017a)	-	9300	DD	Used lexical features and reduced feature sets to classify depressed tweets.
Mowery et al. (2017b)	-	9300	DD	Analysis of social media content should take into account the context in which certain terms are used.
Nguyen et al. (2017b)	$3,221^{(3)}$	769M	LS	Suggested textual and temporal kernel-based features for population health indices prediction.
Yazdavar et al. (2017)	7,046	21M	DD	Guided approach to combine semantically PHQ-related terms in the same topical LDA cluster.
Chen et al. (2018a)	7,968	11.9M	MI	Used SVM and RF to classify 4 types of mental disorders and control groups.
Chen et al. (2018b)	1,185	2.3M	SI	Predict users at risk of depression using temporal analysis of eight Ekman's basic emotions as features.
Coppersmith et al. (2018)	836	395,230	SI	Used GloVe, bidirectional LSTM, and attention mechanism to fetch the most informative terms.

(3) states

References for ML algorithms applied over data collected from Twitter platform by year (Cont.)

Ref.	Users	Posts	Field	Objective
Johnson Vioulès et al. (2018)	60	5,446	SI	Implemented pointwise mutual information measure to detect sudden emotional changes for monitoring suicide warning signs.
Joshi et al. (2018)	200	1.2M	DD	Sentiment, emotion and behavioral features using ensemble classifiers with accuracy of 90.
Nguyen et al. (2019)	$3,221^{(4)}$	1.1B	MI	Estimate population health indices of the US counties using graph-based model.
Weerasinghe et al. (2019)	654	-	DD	After removing direct mention of depression, using SVM with sLDA, BoW, word clusters and POS features achieved Average Precision (0.87).
Weerasinghe et al. (2019)	492	-	PT	Removed direct mention of PTSD, using SVM with sLDA and BoW features achieved average precision (0.88).
Li et al. (2020)	1.4M	80M	DD	CorEx topic modeling with lexicons derived from PHQ-9 to estimate stress related to COVID-19 pandemic in the US.
Razak et al. (2020)	20	-	DD	Online sentiment analysis on the users and followers tweets using VADER, TextBlob and CNN.
Roy et al. (2020)	2,938	4M	SI	RF classifier based on NNs psychological SI indicators and sentiment polarity. The model was validated using SI algorithm estimation on Twitter data and county-wide suicide death rates over 16 days in August and October, 2019.
Verma et al. (2020)	-	15,000	DD	Hybrid deep learning model to predict depressed tweets.

(4) states

Reddit

Reddit is an open-source platform that allows users to publish, comment, or vote on submissions. Reddit had over 430 million active monthly users who collectively have generated 199 million posts with more than 130 thousand active communities, 1.7 billion comments, and 32 billion upvotes during 2019[6]. Posts are grouped by areas of interest that cover a variety of topics, such as gaming, sport, news, and many others. Every subreddit has its own rules, administrators, and subscribers. Some of these subreddit address mental health issues, including anxiety, depression, and suicide. Subscribers may share personal experiences, seek help, and offer support to others. In March 2020, there were nearly 190k subscribers to the *"SuicideWatch"* subreddit and nearly 600k subscribers to the *Depression* subreddit.

Yates et al. (2017) created an experimental dataset that contains around 9k diagnosed users and over 100k control users and named it "Reddit Self-reported Depression Diagnosis (RSDD) dataset". Similarly, Losada and Crestani (2016) collected 481,873 control and 49,580 with signs of depression. Thorstad and Wolff (2019) collected 56,009 posts for each of the following clinical subreddits: r/ADHD, r/Anxiety, r/Bipolar, and r/Depression and used unigram word vectors for training a logistic regression model to classify the post to one of the four mentioned mental illness. The depression classification model achieved an F1-score of 0.74. They also concluded that the language used in a non-clinical context would be predictive of which clinical subreddit the user would later post to.

Table 3.3 presents a summary of recent research done using the Reddit website as a data source. One of the datasets that are used in this research is Reddit-based thereupon it will be illustrated in Section 3.4.4 in more details.

Weibo

Weibo is China's biggest microblog-site, formerly known as "Sina Weibo." At Q3 2020, it has more than 523 million monthly users and more than 100 million user-posted messages per day[7]. There are many studies done using Weibo as a data source. Wang et al. (2018)

[6]https://www.redditinc.com/

[7]https://www.statista.com/statistics/795303/china-mau-of-sina-weibo/#statisticContainer

Table 3.3. References for ML training algorithms applied over data collected from the Reddit platform

Ref	Posts	Notes
Balani and De Choudhury (2015)	32,509	Data-driven approaches for predicting self-disclosure of social media content related to mental health.
De Choudhury et al. (2016)	63,485	Predictive models of online communities to prevent suicidal disclosures.
Gkotsis et al. (2017)	90,518	Adopted NN recognized mental-illness related posts with 91.08% accuracy.
Gkotsis et al. (2017)	197,436	Classified to one of 11 themes disorders with 71.37% accuracy(1).
Kavuluru et al. (2016)	11,730	Designed a system for defining the helpfulness of comments on Reddit forums for mental health.
Alambo et al. (2019)	4,992	Framework for semantic clustering and sequence-to-sequence models to assess the level of suicide risk by question answering mechanism of C-SSRS questionnaire(2).
Thorstad and Wolff (2019)	515,374	Classified posts to one of 4 mental illness subreddits and used non-clinical subreddits to early predict mental illness.
Tadesse et al. (2020)	7,201	Using LSTM-CNN over Word2vec hybrid model discovered shift in language usage of at-risk users.

(1) BPD, bipolar, schizophrenia, anxiety, depression, selfharm, suicideWatch, addiction, cripplingalcoholism, opiates and autism.

(2) Columbia Suicide Severity Rating Scale.

randomly crawled 1 million users (394 million postings) and used a keyword-based method to pinpoint prospect users at risk of suicide. Afterwards, three mental health researchers were assigned to label at-risk users manually. They identified 114 users (60,839 posts) with suicide ideation and used linguistic analysis to explore behavioral and demographic characteristics. Lv et al. (2015) used Weibo to build a suicide dictionary by initial post-sampling and words selection, then irrelevant words elimination, and finally word expansion using Word2vec vector representations. They found that the dictionary-based recognition correlates well with the expert ratings ($r = 0.507$) in detecting suicidal expression along with evaluating the level of suicide risk. Similarly, the association between suicidal ideation and the linguistic features was examined (Cheng et al., 2017; Huang et al., 2014). Whereas, Hao et al. (2013) used Support Vector Machine (SVM) and Neural Networks (NN) on psychological measurement data (SCL-90-R) along with Weibo blogs to identify users with mental health issues. On post-level, Gao et al. (2017) extracted new features based on content and emotions by examining the semantic relationships between the words in a labeled data set of 9,123 microblogs with suicidal ideation.

3.3 Shared Datasets

Datasets can be made available for shared tasks or competitions, such as the CLPsych Shared Task, eRisk and Crisis Text Line datasets that are described in the next subsections.

3.3.1 CLPsych Shared Tasks

In 2014, the workshop on Computational Linguistics and Clinical Psychology (CLPsych) began a collaboration between clinical psychologists and computer scientists and developed links across the research community. The workshop series aims to expedite the development of language technology for mental healthcare with an emphasis on using social media to predict population mental health. Shared tasks provide gold standards because they are built on the same dataset to test and compare various solutions to the same problem under study.

At the 2015 CLPsych workshop, participants were asked to determine whether a user has PTSD or depression, or none of them based on self-reported Twitter diagnoses (Cop-

persmith et al., 2014a). The dataset is composed of 1,146 users: 246 users with PTSD, 327 depressed users and 573 control users that match the age and sex of the former two groups. For all three tasks, the system of Resnik et al. (2015a) performed the best, obtaining an average precision above 0.80. The maximum precision was achieved for the task that distinguishes PTSD vs. control users by training an SVM classifier with a linear kernel based on topic modeling and lexical tf-idf features. Later, Orabi et al. (2018) used optimized word embeddings with a deep learning model and achieved a precision of 87.4% and an F1-score of 86.97%. The CLPsych 2015 dataset is used in this research, thus it will be explained in Section 3.4.

The CLPsych 2016 and 2017 shared tasks invited participants to automatically triage posts collected from the ReachOut.com forum as green, amber, red or crisis to assist the forum moderators to identify and address pertinent cases as soon as possible. A total of 15 teams participated in the task with 60 different submissions. At first, 947 annotated posts were given to each team to develop and train their models. The best-performing system used an ensemble classification approach with tf-idf weighted unigrams and post embeddings, and achieved an F1-score of 0.42 (Kim et al., 2016). Following, Cohan et al. (2017) used the same dataset and applied an ensemble of lexical, LIWC, emotions, contextual and topic modeling features using an SVM model to reach a better F1-score of 0.51.

In 2019, the shared task consisted of three tasks; Task A was about risk assessment of the users who posted in the SuicideWatch subreddit, into one of the following four levels of risk: No risk, low, moderate and high. Task B was about risk assessment using all the subreddits. Task C was about screening users for probabilistic risk from non-mental health-related subreddits. The dataset consists of 1,242 users (including both positive examples and controls). The participation of Mohammadi et al. (2019) under the team name CLaC obtained the best macro-F1 score for task A (0.533) by adding the SVM-predicted class probabilities at the end of the pipeline on top of a set of CNN, Bi-LSTM, Bi-RNN and Bi-GRU neural networks. For Task B, Matero et al. (2019) achieved the best F1- score (0.504) using BERT features extracted separately from SuicideWatch and non SuicideWatch posts. For Task C, a stacked parallel CNN with LIWC and a universal sentence encoder Cer et al. (2018), produced the best unofficial F1 score (0.278) as compared to (0.268) for the CLaC primary system. Finally, Howard et al. (2020) used lexicon analysis, LIWC, Empath,

Word Count, VADER, and DeepMoji for emotional feature extraction to train the model on CLPsych 2017 dataset and test it on CLPSych 2019 expert-labeled dataset with a maximum F1-score (0.616). The CLPSych 2019 dataset is used in this research as well, therefore, more details are presented in Section 3.4.4.

3.3.2 eRisk Shared Tasks

The 2017 CLEF eRisk was a pilot project that extends the CLEF initiatives that have been operating since 2000, leading to the systemic evaluation of information systems, mainly through experiments on shared tasks. The primary purpose of CLEF eRisk 2017 and 2018 (Task 1) was to address issues related to assessment criteria, effectiveness indicators and other early detection of depression mechanisms (Losada et al., 2019a, 2017, 2018). The shared task focused on the automatic detection of the risk of depression from Reddit posts of a user as soon as possible.

For eRisk 2017, the training set has been manually annotated by experts and contains 486 users (83 depressed with 30,851 posts and 403 non-depressed with 264,172 posts). The test set held 401 users (52 depressed with 18,706 posts, and 349 non-depressed wrote 217,665 posts). A total of 30 systems were submitted by eight teams in the pilot task (Almeida et al., 2017). The highest precision was 0.69 submitted by the Biomedical Computer Science Group from the University of Applied Sciences and Arts Dortmund (FHDO) while the highest recall was 0.79 submitted by The LIDIC Research Group, from Universidad Nacional de San Luis. They examined multiple document representations like Bag of Words (BoW), Concise Semantic Analysis, Character 3-grams, and LIWC and using Random Forests, NB, and decision trees machine learning algorithms. The evaluation measures include an early risk detection measure (ERDE) along with standard classification measures, such as F1, Precision and Recall. This measure is mainly to reward correct classification using a fewer number of user submissions and to penalize late decisions.

The 2018 eRisk task continued with the goal of early detection of symptoms of depression, along with a new task of early detection of anorexia indicators. 2017 dataset was used as training set, then additional 820 users were added with more than 500,000 posts for testing. There were 45 contributions from 11 teams. No significant improvement was noticed, and most participants ignored the trade-off between early detection and accuracy.

Consequently, Leiva et al. (2017) used tf-idf features to compare different ML algorithms on the same dataset and concluded that the Random Forest shows the highest precision score while the K-Nearest Neighbours obtains the highest recall. The combination of all of them in the Voting Algorithm present an improvement in the F1 measure.

There were three tasks for eRisk 2019. The first task continued in the same direction as previous challenges for early detection of depression symptoms, the second task is relating to instantly completing a user interaction-based depression questionnaire in social media, and a similar task was introduced for unsupervised self-harm detection. The findings indicate that it is uncertain whether early signs of self-harm can be identified from social media user experiences until they join a self-harm community (Losada et al., 2019b).

As a continuation of the previous tasks, eRisk 2020[8] followed the same line and resumed with two tasks: the early detection of self-harm signs and measuring the severity of depression signs with a bigger training dataset.

3.3.3 Crisis Text Line

The Crisis Text Line[9] supported by Kids Help Phone is a free 24/7 crisis support texting hotline to assist people with mental health issues through texting (Dinakar et al., 2014). As of October 2019, Crisis Text Line has processed more than 100 million text messages. The data is used to study different mental illness trends across the US, including but not limited to depression, self-harm, and suicidal ideation. The result of the analysis is displayed publicly on CrisisTrends.org. Althoff et al. (2016) experimented with roughly 15k counselor messages to evaluate the linguistic aspects of effective therapy. Unigram and bigram features on a regression model with L1 regularization and 10-fold cross-validation showed the best performing model to predict the effectiveness of a patient-counselor conversation with an accuracy of 0.687, and AUC of 0.716. To gain access to an anonymized version of the dataset, the researcher must apply and got accepted in the Research Fellows program sponsored by Crisis Text Line.

[8] https://early.irlab.org/

[9] www.crisistextline.ca

3.4 Datasets Used in This Research

For population depression and suicide prediction, different datasets are used to train our models. Table 3.4 summarizes the list of all the datasets used in this research, followed by a description for each.

Table 3.4. Summary of the datasets used in this study

Topic	ID	Platform	Users	Posts	Ref.
Population	$\mathcal{P}1$	Twitter	40,631	2,591,108	ASI-2015
Depression	$\mathcal{D}1$	Twitter	897	1,533,626	(Coppersmith et al., 2014a)
	$\mathcal{D}2$	Twitter	3,536	1,690,955	(Shen et al., 2017)
Suicide	$\mathcal{S}1$	Reddit	621	1,105	(Zirikly et al., 2019)
	$\mathcal{S}2$	Twitter	495	987,870	ASI Annotation

3.4.1 Advanced Symbolics Inc. Dataset

The $\mathcal{P}1$ and $\mathcal{S}2$ are subsets from the dataset that is collected by Advanced Symbolics Inc. (ASI), a market research company based in Ottawa, Canada[10]. ASI is maintaining a subset of tweets posted by Twitter users that is statistically representative of Canada's population. By 2018, they have collected millions of tweets for 278,627 users. They use the Conditional Independence Coupler (CIC) sampling algorithm that is based on Coupling from the Past (CFTP) with enhancing the stopping condition by measuring how far is the chosen node (user) from the starting node on a smaller subset of the online network (White et al., 2012). The method adjusts the weights of sampling using post-stratification to compensate for the underrepresented groups of the population. The algorithm is mathematically proven to generate a representative sample of the population and the author verified the representativeness of the sample by comparing 3,032 Toronto Twitter user profiles using their tweets during 2011 with the same year's census patterns, the Twitter users demographics followed the demographics of 2011 census (White, 2016). $\mathcal{P}1$ dataset contains population representative Canadian users who tweeted during 2015, whereas the $\mathcal{S}2$ dataset will be explained in more details in Section 3.4.5.

[10]https://advancedsymbolics.com/

Both datasets contain the following information per user:

- **Spatial information**: The location of Twitter users can be inferred either by using the GPS coordinates of the tweets - if enabled by the user from his/her mobile device - or by the self-declared location - set by the user in his/her profile page. If the location property is enabled, then the longitude and latitude points are fetched from the user's mobile GPS and stored with each tweet till it is turned off. The location of the user is predicted using the k-means algorithm to cluster the GPS coordinates, and presumes that the cluster of the greatest number of points as the user's location. If no geotagged tweets exist, then Microsoft's Bing Maps is used to look for the address specified in the user's profile (if feasible); otherwise the inferred-location field is left empty. The CIC algorithm can limit the searching within users in a certain geographic area such as Canada or Montréal. Table 3.5 shows the description of the recorded spacial fields.

Table 3.5. Spacial information data dictionary

Field	Level	Type	Description
Geotag	Tweet	Coordinate	When a user shares his/her location at tweet time, the longitude and latitude coordinates are stored as the exact GPS location at Tweet time.
Location	User	String	The location field contains user's specified location in the user's public profile.
Province	User	String	Based on the location field, the province value is set based on Canada's city/province mapping table.

- **Demographic information**:

  ASI predicts users' demographics using signals from Twitter data and other related external resources. The age and sex probability distribution is estimated by analyzing the profile photo using Face++ (a deep learning system developed by Megvii Technology to obtain face attributes including age and sex). Then the probability distribution is adjusted by comparing the first name with Canada's birth records and the life tables[11] that contains life expectancy and associated age and sex projections for Canada (Daneshvar et al., 2019).

[11] https://www150.statcan.gc.ca/n1/en/catalogue/84-537-X

It should be pointed out that sex and gender are two different terms. Social statistics systems are more likely to use the *Gender Classification*, whereas demographic and health indicators usually use the *Sex Classification* based on sex at birth[12]. The age and sex probability distribution is deduced for each user in 12 fields as follows:

$$\mathbb{A} = \{\text{“} < 25\text{”}, \text{“}25 - 34\text{”}, \text{“}35 - 44\text{”}, \text{“}45 - 54\text{”}, \text{“}55 - 64\text{”}, \text{“} > 65\text{”}\}$$

$$\mathbb{S} = \{\text{“}Male\text{”}, \text{“}Female\text{”}\}$$

$$Sex_{age} = \{P(s_a) : \forall s \in \mathbb{S} \wedge a \in \mathbb{A}\}$$

$$\text{where} \sum_{(\forall s \in \mathbb{S})} P(s) = 1 \text{ and } \sum_{(\forall a \in \mathbb{A})} P(a) = 1$$

The differences between the probabilities of each category vary. Thus, we decided to keep users with high confidence for both *Age* and *Sex* prediction based on the following rules:

- Sex: We assign to the user the sex of the maximum sex probability of all age groups with a probability more than 92.5% , using the following equation:

$$P(\mathcal{S}) = \{\max(P_{Male}, P_{Female}) : |P_{Male} - P_{Female}| > \epsilon; \epsilon = 0.85\}$$

- Age: We assign to the user the age group of the maximum age group probability ($P(\alpha)$), given that the difference between the largest and the second largest is greater than ϵ[13], using the following steps:

$$P(\alpha) \leftarrow \max\{\sum_{(\forall \zeta \in \mathbb{S})} P(\alpha) : \alpha \in \mathbb{A}\}$$

$$\mathbb{B} \leftarrow \mathbb{A} - Age(P(\alpha))$$

$$P(\beta) \leftarrow \max\{\sum_{(\forall \zeta \in \mathbb{S})} P(\beta) : \beta \in \mathbb{B}\}$$

$$\text{and} |P(\alpha) - P(\beta)| > \epsilon$$

[12]Statistics Canada, recommended the following definition on January 25, 2018: "Gender refers to the gender that a person internally feels ('gender identity' along the gender spectrum) and/or the gender a person publicly expresses ('gender expression') in their daily life, including at work, while shopping or accessing other services, in their housing environment or in the broader community. A person's current gender may differ from the sex a person was assigned at birth (male or female) and may differ from what is indicated on their current legal documents. A person's gender may change over time". https://www23.statcan.gc.ca/imdb/p3Var.pl?Function=DEC&Id=410445

[13]We choose $\epsilon = 0.35$, since each age group has a probability of 0.17

- Race: The available race attributes in ASI dataset represents the race predicted based on the US census, i.e. *Asian, Black, Hispanic* and *White*. This attribute is not considered in this research.

- **Temporal information**:
  - User Creation time and date: This field contains the UTC datetime that the user account was created on Twitter, and this field is not used in this research.
  - Tweet time and date: This field contains the UTC date and time when the user posted the tweet. This field can give insights on the usual messaging time and may indicate sleeping patterns.

- **Textual information**:
  - User description: The text written by Twitter users explaining their profiles.
  - Tweet text: A short status update posted by the user with a limit of 140 characters, and this is the most crucial field in this research. In 2017, Twitter doubled the character count to 280, but that did not change the tweeting attitude of most users. As reported by Twitter, only 5% of the Tweets use more than 140 characters[14].

After removing the records of the users who do not have a well-defined location mapped to a Canadian province, and the users with age or sex prediction less than the defined confidence, the number of users reduced from 278,627 to 40,631. On the post level, we also filtered out posts with less than 32 characters, that decreased the number of posts from 9,304,441 to 2,591,108 tweets.

Canada has ten provinces and three territories. The distribution differences based on sex between the 2016 Canadian census and the Twitter users in $\mathcal{P}1$ who tweeted during the same period are <5% difference for each area as shown in Table 3.6, except for the province of Quebec. Quebec is the only province whose sole official language is French[15]. Less than 10% of Quebecers are Anglophone[16]. The language of correspondence considered in this research is English and for that reason, the province of Quebec is excluded.

[14]https://blog.twitter.com/en_us/topics/product/2017/tweetingmadeeasier.html
[15]https://en.wikipedia.org/wiki/Official_Language_Act_(Quebec)
[16]https://en.wikipedia.org/wiki/Language_demographics_of_Quebec

Table 3.6. Geographic population difference between the Canadian census 2016 data and $\mathcal{P}1$ dataset

Pr.	**Province**	$\mathcal{P}1$	**Census Data**	**Diff.%**
AB	Alberta	5,311	3,279,397	-1.6%
BC	British Columbia	7,309	3,943,043	-4.3%
MN	Manitoba	1,493	1,011,492	-0.2%
NB	New Brunswick	805	617,041	0.2%
NL	Newfoundland & Labrador	690	434,203	-0.2%
NT	Northwest Territories	191	35,054	-0.3%
NS	Nova Scotia	2,075	765,190	-2.4%
ON	Ontario	17,150	11,042,167	-3.7%
PE	Prince Edward Island	454	117,350	-0.7%
QC	Quebec	3,681	6,561,788	13.8%
SK	Saskatchewan	1,393	866,267	-0.4%
YT	Yukon	79	31,411	-0.1%

Table 3.7. Population difference between the 2016 census data and the $\mathcal{P}1$ dataset based on Sex (+)Female / (-)Male

Pr.	$\mathcal{P}1_M$	$\mathcal{P}1_F$	$\mathcal{C}_{\mathcal{M}}$	$\mathcal{C}_{\mathcal{F}}$	Diff.%
AB	2,480	2,831	1,615,161	1,664,236	2.60
BC	3,533	3,776	1,986,602	1,956,441	2.00
MN	687	806	503,360	508,132	3.70
NB	370	435	309,332	307,709	4.20
NL	366	324	218,374	215,829	-2.80
NS	87	104	16,878	18,176	2.60
NT	1,060	1,015	389,682	375,508	-0.20
ON	7,893	9,257	5,581,149	5,461,018	4.50
PE	231	223	59,866	57,484	0.10
SK	574	819	425,094	441,173	7.90
YT	29	50	15,488	15,923	12.60

Regarding the sex distributions in provinces, the differences between the population share of the two sexes in the 2016 census and $\mathcal{P}1$ dataset have differences between [0 - 5] as shown in Table 3.7, except the province of Saskatchewan and in the Yukon territory. The male population is under-represented in both areas and this needs to be considered when analysing our results. Nunavut users were removed based on the age and sex defined restrictions. Nunavut Territory, has a population of 35,944; i.e., less than 0.1% of the Canadian population, 32.5% of them less than 15 years and 63.1% have Inuktitut as their mother tongue language.

3.4.2 CLPsych Shared Task 2015 Dataset

$\mathcal{D}1$ is Twitter-based dataset collected based on publicly self-report diagnosis of depression created for CLPsych shared task 2015 (Coppersmith et al., 2015b). The shared task focused on three binary classification tasks to identify depression, PTSD, and control users. Since this study is concerned with major depressive disorder, not post-traumatic stress disorder, we omit the PTSD users and focus on only one of the sub-tasks, which identifies depression users versus control users. We use only the training dataset that consists of 327 depression users and 572 control users because of the absence of the test dataset labels.

For each user, the most recent 3,000 public tweets were collected with a minimum of 25 tweets. The tweet of the original diagnostic statement was removed from the dataset. The tweets collected were posted between 2011 and 2014. Table 3.8 summarizes the overall statistics of $\mathcal{D}1$ dataset.

Screen names were anonymized systematically to preserve the consistency of the references throughout all tweets. The user information contains the number of friends, followers, favorites, the time the account was created, and the user's time zone along with the estimated age and sex. Figure 3.2 shows the age and sex distribution of $\mathcal{D}1$ dataset users. Only 0.6% of the users are of age ≥ 40 and most of the users (46.9%) are males between the age of 19 and 29.

The post information contains the tweet text, the timestamp, and the language identification of the text. All fields that may lead to revealing the user's identity were removed, including geolocation information. Also, the mentioned URLs were masked beyond the domain name.

Table 3.8. $\mathcal{D}1$ dataset statistics for control and depression users

Description	Control	Depression
Number of users	572	325
Average age	24.35	21.70
% female	74.3%	69.9%
Total number of posts	964,748	568,878
Average number of posts	1,686.6	1,750.4
Total posts words	10,180,701	6,407,988
Average posts words	17,798.4	19,716.9
Total posts chars	65,661,468	39,681,782
Average posts chars	114,792.8	122,097.8
Total emoticons	105,685	66,365
Average emoticons	184.8	204.2

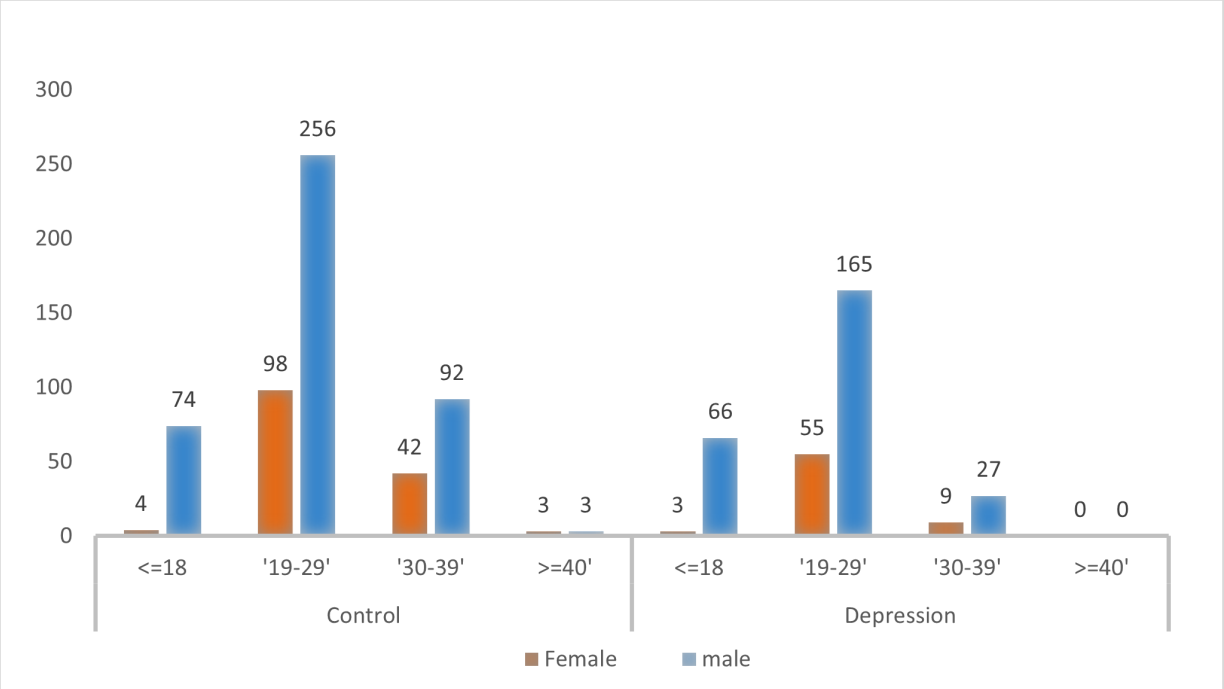

Figure 3.2. Distribution of age and sex for the users of $\mathcal{D}1$ dataset

3.4.3 Depression Self Reported Dataset

Shen et al. (2017) followed the same guidelines of Coppersmith et al. (2014a) and collected 1,402 depressed users who posted 292,564 tweets within one month against 5,610 non-depressed users who posted 3,953,183 tweets between 2009 and 2016. A user is labeled as depressed if there exist a tweet that indicates explicitly that he/she was diagnosed with depression by fulfilling the exact pattern *"(I'm/ I was/ I am/ I've been) diagnosed depression"*[17]. Tweets posted within a month of the anchor tweet were collected. The control users are labeled as non-depressed if the user never posted any tweet containing the character string *"depress"*. This dataset contains all the fields returned from Twitter API in json files format. To be consistent with all the other datasets, we kept the tweets text accompanying its timestamp. The user profile information does not contain the demographics of the user. Thus, we used M3-inference tool to predict the age and sex of the user (Wang et al., 2019b). M3 is a multi-learning, multilingual and multi-attribute deep learning system that infers demographic attributes from social media profiles. M3 was trained on a huge dataset of tweets, an image dataset of faces from IMDB and a Wikipedia and crowdsourced dataset for all the attributes. It also classifies the account as organization or individual account. From the same source (Shen et al., 2017), an unlabeled depression candidate dataset was added where users were obtained if their tweets containing the *"depress"* character string resulting 6,672 users in total. After age-sex prediction -based on M3 - we filtered users with CI=95.0 for sex and organization categories. We assigned the max_age category for each user based on the following four age categories $\{<= 18, 19-29, 30-39, >= 40\}$, the age distribution in shown in Table 3.9. The sex and age distribution is illustrated in Figure 3.9. The final dataset contains 3,536 users and the statistics of this dataset is reported in Table 3.10.

Table 3.9. Age_group distribution for $\mathcal{D}2$ dataset

Label/Age Group	<= 18	19-29	30-39	>= 40
Control	0.240	0.427	0.164	0.169
Depression	0.226	0.391	0.208	0.175

[17]We suggest that additional manual or automatic annotation should be performed to filter out ingenuine declaration such as sarcasm, quotation,..etc.

Table 3.10. $\mathcal{D}2$ dataset statistics for control and depression users

Description	Control	Depression
Number of users	2,011	1,525
% female	42.6%	65.6%
Total number of posts	1,404,110	286,845
Average number of posts	698.2	188.1
Total post words	15,053,584	3,589,727
Average post words	7,485.6	2,353.9
Total post chars	98,575,481	22,856,405
Average post chars	49,018.1	14,987.8
Total emoticon	47,559	34,287
Average emoticon	23.6	22.5

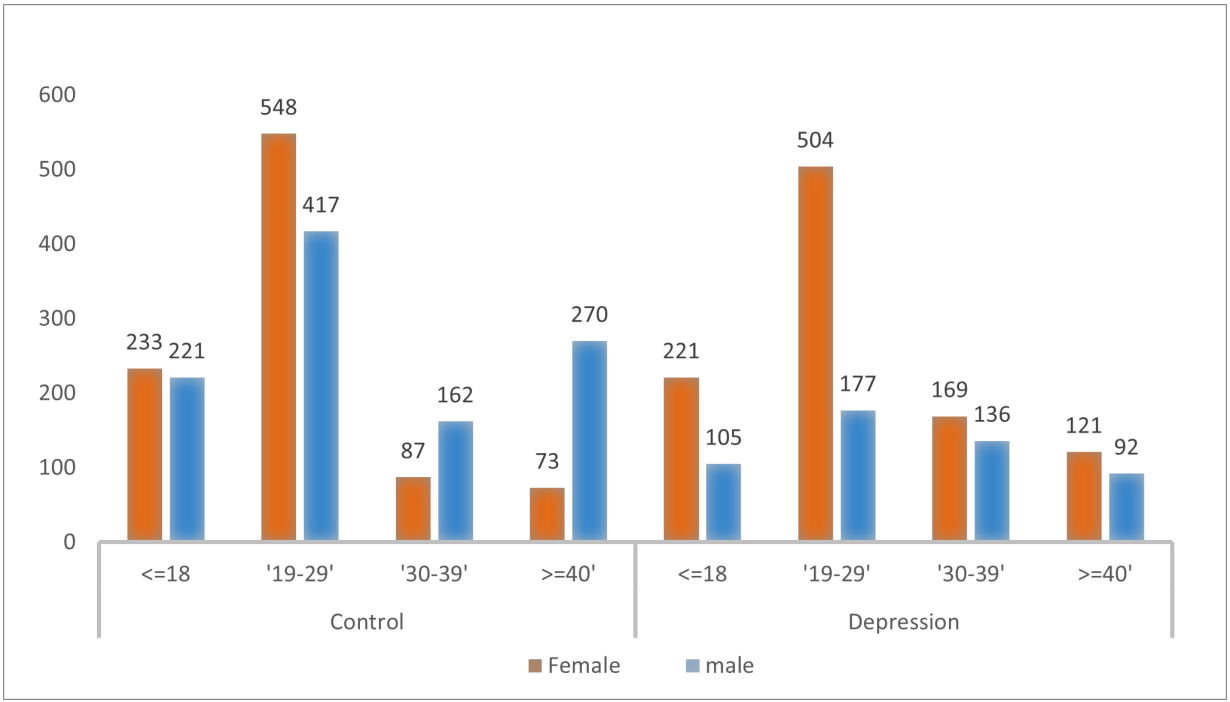

Figure 3.3. Distribution of age and sex for the users of $\mathcal{D}2$ dataset

3.4.4 CLPsych Shared Task 2019 Dataset

The $\mathcal{S}1$ dataset is based on the CLPsych 2019 shared task (A) (Zirikly et al., 2019). The task (A) dataset consists of the posts of users who posted only on the SuicideWatch subreddit between 2006 and 2015. The severity of the users' status was annotated by crowdsourcers into four classes: *non-suicidal, low-risk, high-risk* and *severe*, with Krippendorff's α= 0.55. For the purpose of this research, we consider the distinction between non-suicidal and suicidal users, regardless of the degree of risk. The $\mathcal{S}1$ dataset contains 909 posts for 462 users with suicide thoughts and 196 posts written by 159 control users as summarized in Table 3.11. $\mathcal{S}1$ dataset was the only well curated dataset available at the time of this research.

The dataset fields contains the *post_id* that identifies the post written by the user, the *subreddit* that is the discussion forum where the post appeared, the *post_title* is the post title optionally entered by the user and the *post_body* is the complete text of the post with a limit of 40,000 characters. The *timestamp* is the date and time of posting using Unix epoch unified time.

The data does not contain any profile information, since measures were taken to protect users' privacy by performing the following steps. First, usernames were replaced with arbitrary numeric identifiers for the *user_id*. Then, automatic processing was performed on post titles and bodies to replace any existing email addresses, URLs, and person entities with $\_EMAIL\_$, $\_URL\_$, and $\_PERSON\_$ identifiers, respectively. Finally, the posts were filtered to the posts that contain English letters only.

Figure 3.4 shows the differences in term usage between suicidal and not at-risk users for $\mathcal{S}1$ using Scattertext. Scattertext is a publicly available tool for visualizing linguistic variation within different categories using scaled F-score (Kessler, 2017). The terms in the figure are the most frequent terms after removing stopwords and words with less than three letters. Terms that are more associated with "Suicidal" are in the blue color, while those connected with "Control" are in the red color. The darkness of the color shows the relevant frequency of the term. The terms that are most characteristic to the user type are displayed on the most right of the figure. Needless to say, that control users use 3rd person singular pronouns apparently in reflexive forms like *himself* and *herself*, while suicidal people use only 1st person singular pronouns like *myself*, with the apparent usage

Table 3.11. $\mathcal{S}1$ dataset statistics for control users and users with the risk of suicide ideation

Description	Control	Suicidal
Number of users	159	462
Total number of posts	196	909
Average number of posts	1.2	2.0
Total title words	2,031	7,109
Average title words	12.8	15.4
Total title chars	10,742	35,896
Average title chars	67.6	77.7
Total body words	27,659	196,914
Average body words	174.0	426.2
Total body chars	145,164	1,018,929
Average body chars	913.0	2,205.5
Total emoticons	24	103
Average emoticons	0.2	0.2

of negative words like *dead, sick, stupid* and *waste*. Also, it is evident for human readers that non-suicidal users are mainly posting in r/SuicideWatch to seek help for others like cousins or friends.

Figure 3.5 shows terms frequencies based on corpus characteristics. That is, the terms that appear regularly in the posts under analysis and considered as descriptive of the corpus as a whole -$\mathcal{S}1$ in this case - but are relatively infrequent compared to general word frequencies.

3.4.5 Suicide ideation ASI Dataset

$\mathcal{S}2$ dataset contains the same attributes of $\mathcal{P}1$ dataset as both are subsets from the ASI collected data as described in Section 3.4.1. However, no records were eliminated based on location or demographics information. The users have a distribution of probabilities for age and sex as shown in Figure 3.6. 67% of the users are estimated in the age between 55 and 64 and only 25% are females. $\mathcal{S}2$ contains 987,870 tweets from 2017 till Sep. 2018 for 495 users. Tweets have been compiled on the basis of phrases such as: *"I want to*

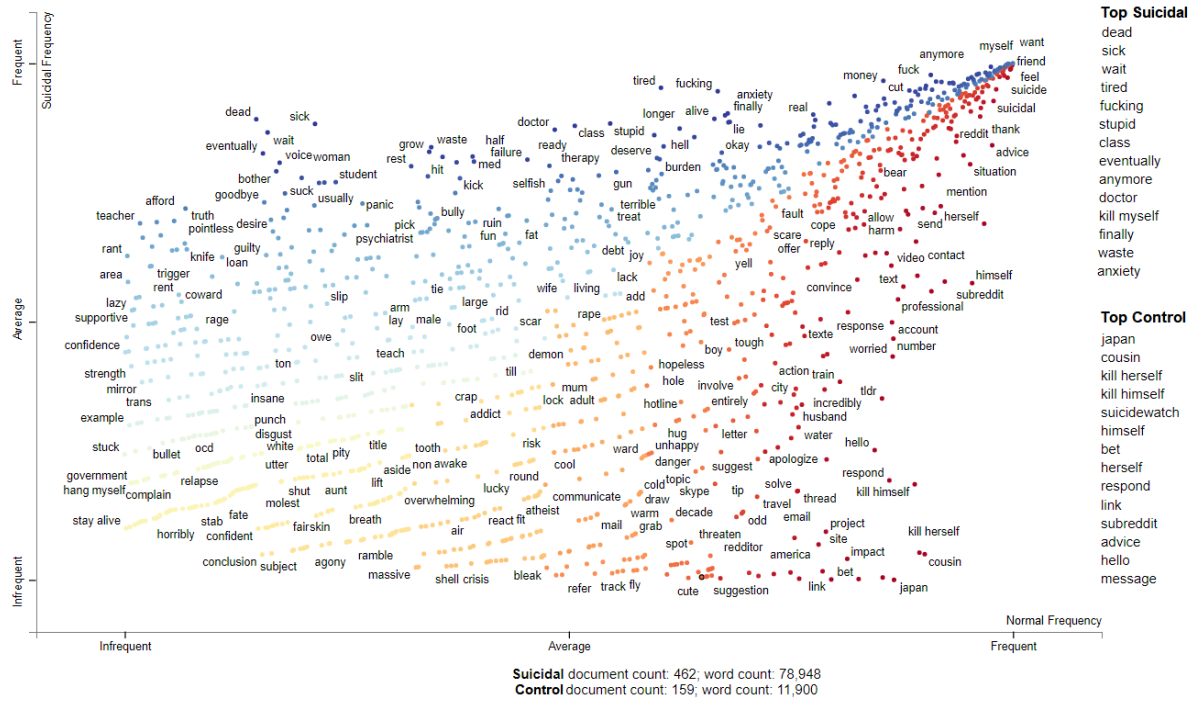

Figure 3.4. Frequency term differences between suicidal and control users for $\mathcal{S}1$ dataset (users who posted in r/SuicideWatch subreddit).

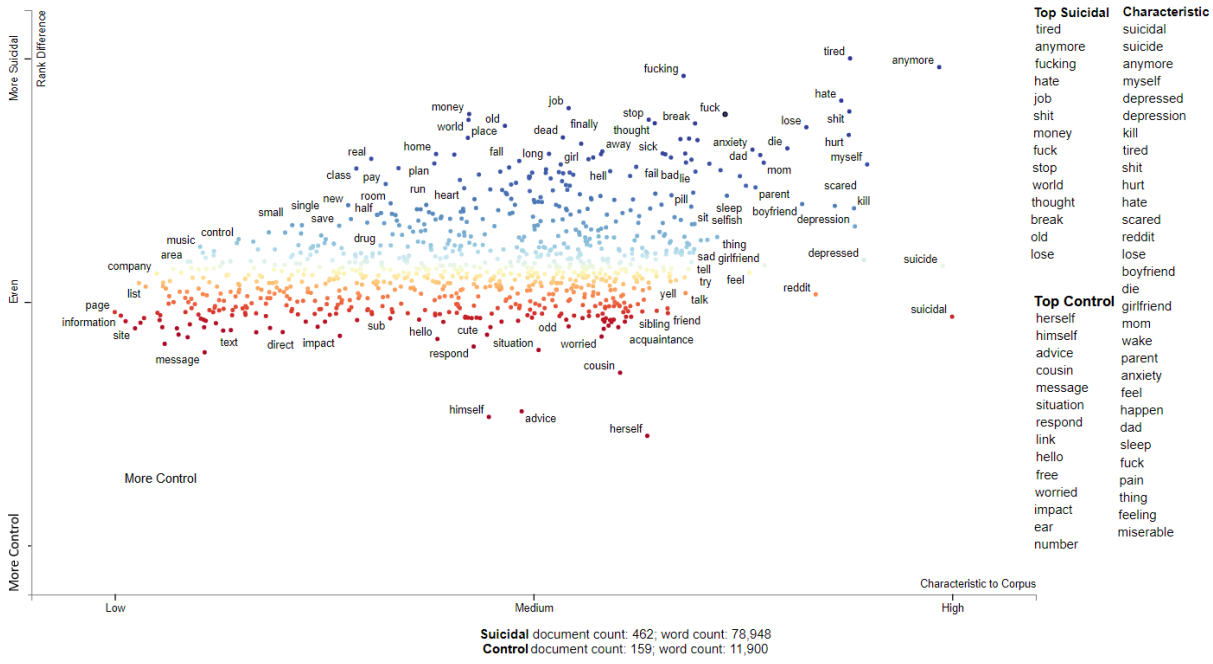

Figure 3.5. Terms frequencies based on $\mathcal{S}1$ corpus characteristics

die", "thinking how to kill myself", "can't do this anymore" etc. Jokes and sarcastic tweets were removed, then tweets were labeled by Registered Psychologists and Clinical Social Workers[18] to at-risk or no risk labels. 654 tweets were labeled as 'at-risk'.

Table 3.12. $\mathcal{S}1$ dataset statistics for tweets collected to measure the risk of suicide based on ASI dataset

Description	No-risk	At-risk
Total number of posts	987,216	654
Total title words	14,496,215	12,538
Average title words	14.68	19.17
Total title chars	93,593,925	69,654
Average title chars	94.81	106.50
Total emoticons	20,447	346
Average emoticons	0.02	0.53

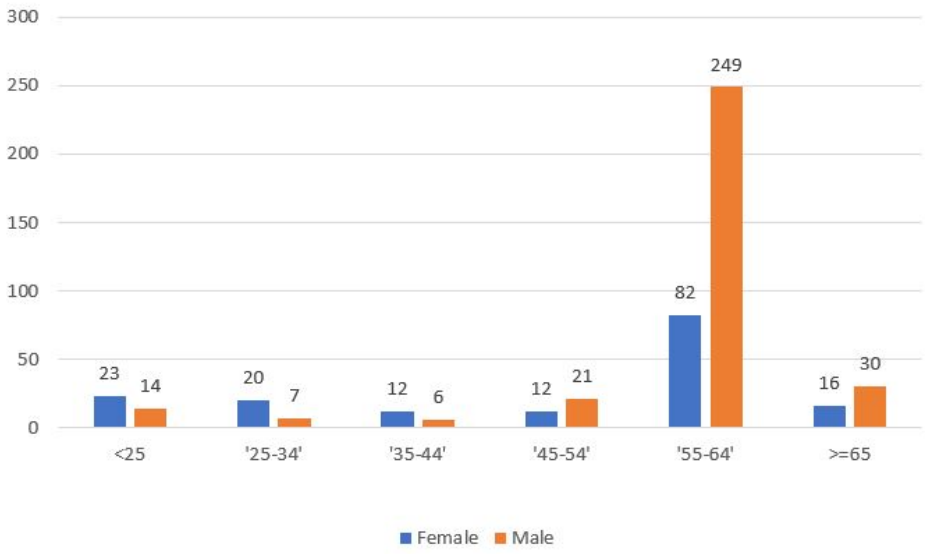

Figure 3.6. Age and sex distribution in $\mathcal{S}2$ dataset

[18]Canadian Association for Suicide Prevention – 30th Annual National Conference (CASP)

3.4.6 How Are These Datasets Being Used?

We train two binary classifiers to discriminate between control group and suicidal or depressed social users media using $\mathcal{S}1$ and $\mathcal{D}1$ respectively , then repeat testing the tuned classifiers on unseen datasets $\mathcal{S}2$ and $\mathcal{D}2$ till we reach a satisfactory predictive model to be able to apply it on the population-level dataset $\mathcal{P}1$, as show in Figure 3.7.

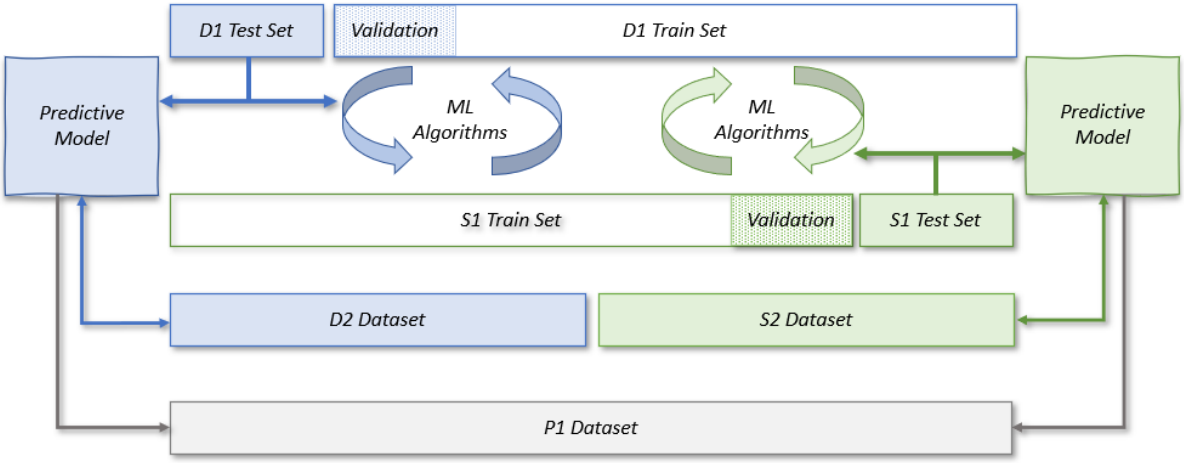

Figure 3.7. Training and testing datasets

3.5 Summary

In this chapter, we introduced different methods for data collection, then emphasized the usage of social media data in detecting signals of mental health problems. Then, five datasets and their attributes have been presented. Finally, we illustrated how these datasets will be used in this research. In the next chapters, we will employ the described datasets to predict depression and suicide ideation for social media users, and then at the population level.

Chapter 4

Depression Prediction from User to Population

4.1 Overview

In this chapter[1], we explore the task of detecting signs of depression from tweets where people often express their feelings, thoughts, interests and opinions implicitly or explicitly. The task we explore is formulated as follows: given a set of tweets that belongs to randomly selected users, the automatic natural language system needs to estimate whether the user is suffering from a major depressive disorder or not. Thus, we propose using different sets of features for classifying Twitter users into depressed or non-depressed users using their textual content. The feature sets used in our models are summarized in Table 4.1 and explained in Section 4.2. The results of the classifiers are discussed in Section 4.3, as Subsection 4.3.1 demonstrates the traditional machine learning results, followed by the results of deep learning models for the same task, in Section 4.3.2. In 4.4, we apply the best model - based on recall and F1-score metrics[2] on test data set - to predict depression within population and compare the results with the depression statistics reported by Statistics Canada for 2015.

[1] Parts of this chapter are published in Skaik and Inkpen (2020a)

[2] $Recall = \frac{TP}{TP+FP}$ and $F1 = \frac{2*TP}{2*TP+FP+FN}$; *TP: True Positive, FP: False Positive, FN: False Negative*

Table 4.1. Feature sets used in traditional machine learning algorithms

Features	**Dimensions**
Statistical	16
LIWC	93
POS tags	45
Tf-idf	200
Topics	50
Sentiments, emotions and major life events	11
Word embeddings	200

4.2 Depression Classification Methods

We use the $\mathcal{D}1$ dataset for evaluating different models. Subsequently, the $\mathcal{D}2$ dataset is employed to test the models' ability to generalize. We aim to apply the best model on $\mathcal{P}1$ dataset to estimate the presence of the depression disorder within the Canadian population and compare the demographics of the predicted depressed users with Canadian statistics for 2015. We selected the $\mathcal{D}1$ dataset (897 users) for training because the approach of annotating the users was based on experts annotations. In addition, the $\mathcal{D}2$ dataset (3536 users) is a Twitter data, and is more suitable for testing the generalizability of the model because of its resemblance to the tweets collected from the population, in terms of variation, scale, and noise. The following sections cover the feature sets, the traditional and deep learning models used in our experiments:

4.2.1 Feature Selection

- **Statistical Features:** Statistical features include the number of words per tweet, word density, number of unique words used, number of characters per tweet, the frequency of posting during different periods of the day: morning (6 a.m.- 12 p.m.), noon (12 p.m. - 6 p.m.), evening (6 p.m. - 12 a.m.) and night (12 a.m. - 6 a.m.), the use of emoticons and emojis in tweets, the usage of different letter case, number of different punctuation, exclamation or question, hastags, mentions and URLs.

- **Tf-idf:** The term frequency (tf) of a document is the normalized occurrence of a term within a document, but most used terms such as *"the"* and *"is"* will have high term frequency with no added value. On the other hand, the inverse document frequency (idf) measures the value that the term provides to the document, for that tf-idf represents the significance of a term in a document based on the frequency of its appearance in the current document and in the other documents in the corpus.

- **Linguistic Features**: Another valuable feature set is LIWC. LIWC defines the language patterns of the posts and categorizes them in psychologically meaningful groups. The output of LIWC tool is a vector of 93 elements capturing different aspects including summary language variables (analytical thinking, clout, authentic and emotional tone), standard linguistic metrics, terms representing psychological perceptions, personal concern categories, informal language indications and punctuation usage. The LIWC 93 features and examples are listed in Appendix A.1.

- **POS-tags:** Part of Speech (POS) tagging is used to capture similar syntactic characteristics. POS tags may lead to improving the performance of the predictive model.

- **Topic Modeling:** Topic modeling is a probabilistic model for finding hidden semantic structures. It is an unsupervised method that considers the set of user posts as an aggregate of latent topics in which a topic is a distribution of co-occurring terms. Different terms which convey an associated aspect are grouped together.

- **Sentiments, Emotions and Major Life Events:** Sentiment analysis was calculated using VADER (Valence Aware Dictionary and sEntiment Reasoner) tool. VADER is a rule-base model used for social media text sentiment analysis. It estimates the sum of all the lexicon ratings and assigns a probability for negative, positive or neutral sentiments. It resembles the annotations of human raters with 0.88 correlation coefficient (Hutto and Gilbert, 2014). We extracted all the emoticons and emojis from the original post and replaced them with their meaning, when possible. We used text2emotion[3] a lexicon-based package to recognize 5 different emotion categories: *Happy, Angry, Sad, Surprise and Fear*. Similarly, we developed a major life events detection for death and divorce, since these life events are as-

[3]https://github.com/aman2656/text2emotion-library

sociated with higher prevalence rates of major depression in both male and female users (Kessler and Bromet, 2013; Burcusa and Iacono, 2007). Major life events are analyzed by manually finding the appropriate words that indicate life events within the pre-processed text that represents the user, then categorize each word to build word lists for each category. Finally, the words found in the text that are relevant for each category are counted. Examples of such lexicons are:
Divorce event = {'divorce', 'separate', 'breakup', 'split',... etc.}
Death event = {'death', 'passed' , 'died', 'sympathy', 'condolences',... etc.}

- **Word Embeddings:** Finally, we exploit the linguistic features extracted using word embeddings. Word embedding represents a document's vocabulary using language modeling and feature learning techniques in natural language processing (NLP) as a dense vector that captures the terms' semantics. Using unsupervised learning approaches, vocabulary terms are initialized with fixed-length continuous-valued vectors then trained using a large corpus of text to calculate distributed representation of words using vector arithmetic based on the company it keeps. The resulting word embedding vectors can be either context-independent such as Word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) or context-dependent such as ElMo (Peters et al., 2018), GPT-3 (Brown et al., 2020) and BERT (Devlin et al., 2018) models.

  Context-dependent word embeddings have different representations of the same term based on its context. On the other hand, Context-independent word embeddings have one vector representation for a term regardless of its different meanings.

  For example, based on GloVe pre-trained vectors on 5.6B uncased tokens and 400K vocabulary from Wikipedia 2014 and Gigaword, the word *"present"* has the following vector representation (using 50-dimensions):

  $\overrightarrow{W}$("present") = array([0.72498 , 0.57224 , -0.24602 , -0.1493 , 0.89771 , 0.66096 , -0.10267 , -0.58517 , -0.49753 , -0.26535 , 0.30384 , -0.16906 , 0.22045 , -0.25927 , 0.40033 , 0.65491 , 0.0078193, -0.13453 , 0.1038 , -0.045302 , 0.23506 , 0.16118 , 0.40373 , -0.2722 , 0.28994 , -0.79136 , -0.33561 , 0.070869 , -0.30566 , 0.10933 , 3.1193 , -0.050093 , -0.32951 , -0.98609 , 0.35644 , -0.17026 , 0.40967 , 0.2018 , -0.24602 , 0.14137 , -0.60586 , -0.14066 , 0.011624 , -0.13846 , -0.84449 , 0.081368 , -0.44129 ,

0.023594 , 0.11587 , -0.30429]

The following are the cosine similarities between the word *"present"* and two of its meanings: *"gift"* and *"the tense"*, it appears that the word vector - more likely - represents the second meaning.

Cosine similarity between 'present' and 'gift'/'surprise: 0.438/0.446
Cosine similarity between 'present' and 'now'/'current' : 0.776/0.775

Word embeddings can be obtained either by using a pre-trained model or by training word vectors on the corpus under study from scratch. There are many word embedding techniques. Word2vec models can be trained with different parameters, including the original continuous bag-of-word (CBOW) and skip-gram (SG) using either hierarchical softmax or negative sampling as illustrated in Figure 4.1.

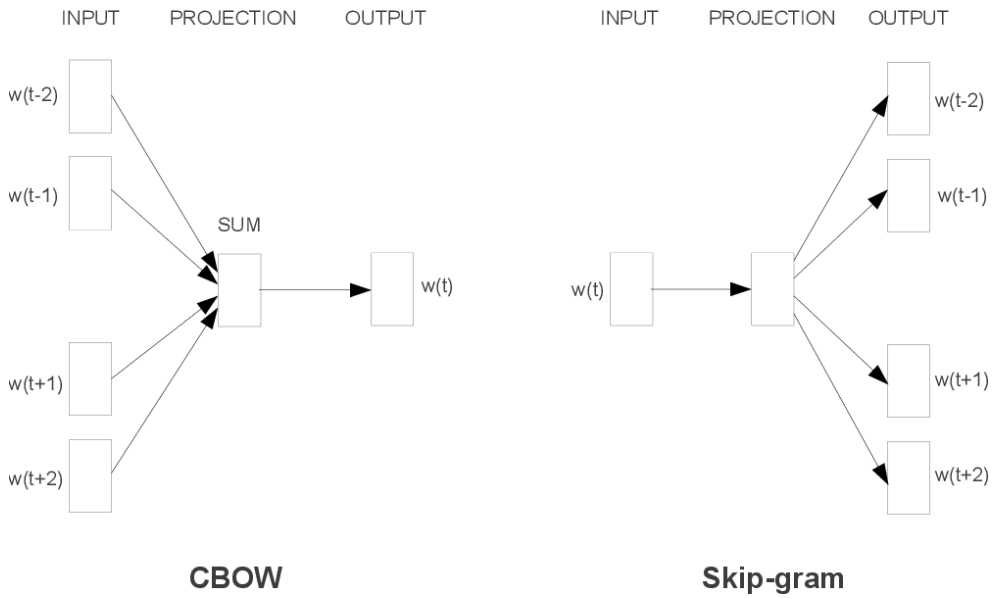

Figure 4.1. (left) CBOW: given a context word w(t), the neural network attempts to predict its context words with a window (k), $w(t-k)...w(t-1), w(t+1)...w(t+k)$ of an input word w(t). (right) Skip-gram: given context words $w(t-k)...w(t-1), w(t+1)...w(t+k)$, neural network attempts to predict the word w(t). In this case, t=2. Image taken from (Mikolov et al., 2013)

Table 4.2 summarizes the word embeddings that we used for depression classification.

Table 4.2. Word embeddings features used for depression detection

Method	Library/Model	Pre-trained	Dimensions
Mean Word Vector	MeanEmbeddingVectorizer	-	200
Tf-idf Mean Word Vector	TfidfEmbeddingVectorizer	-	200
GloVe-Twitter-2B tweets	27B tokens, 1.2M uncased vocab	✓	200
GloVe-Common	840B tokens, 2.2M cased vocab	✓	300
Document Embedding	Doc2Vec	-	200
Numberbatch	ConceptNet	✓	300
fastText-Wiki	16B tokens,1M vocab	✓	300
fastText-Crawl	600B tokens, 2M vocab	✓	300

4.2.2 Traditional Classification Models

We experimented with several traditional machine learning algorithms for detecting depressed users based on their tweets. SVM and LR are discriminative models that are wildly used for detecting health-related tweets (Paul and Dredze, 2017). SVM employs kernel functions to represent each instance as a point in a higher dimensional space then finds the optimal hyperplane that maximizes the margin between the classes. Whereas LR is used to estimate the likelihood of occurrence of one dependent binary variable based on one or more independent variables by fitting the data to the logistic curve.

Tree ensemble methods use boosting or bagging techniques to consolidate several decision trees for a better predictive system than employing a single decision tree. Random Forest (RF) is an extension over bagging. Each decision tree in the forest uses a random set of features and trains its classifier based on a random subset of the training dataset. Gradient Boosting is an extension over boosting method and GBDT constructs an additive model step-by-step; each time the regression tree is fit on the negative gradient of the binomial loss function[4].

[4]https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

4.2.3 Deep Learning Classification Models

Deep Learning algorithms have obtained superior results on multiple language-related tasks in comparison to conventional machine learning algorithms. And hence the target is to increase the recall and F1-score in the test dataset, we conducted more experiments using various deep learning models. Following are the different architectures of neural network that we used in this research:

- CNN

  Convolutional Neural Networks (CNN) is one of the most common deep learning approaches for text classification. CNN is most widely used in image processing applications such as object detection, image classification and facial recognition. However, more recently CNN started to be used more and more to solve NLP tasks such as machine translation, emotion and sentiment analysis, document summarization and simplification, etc. A classification example is shown in in Figure 4.2.

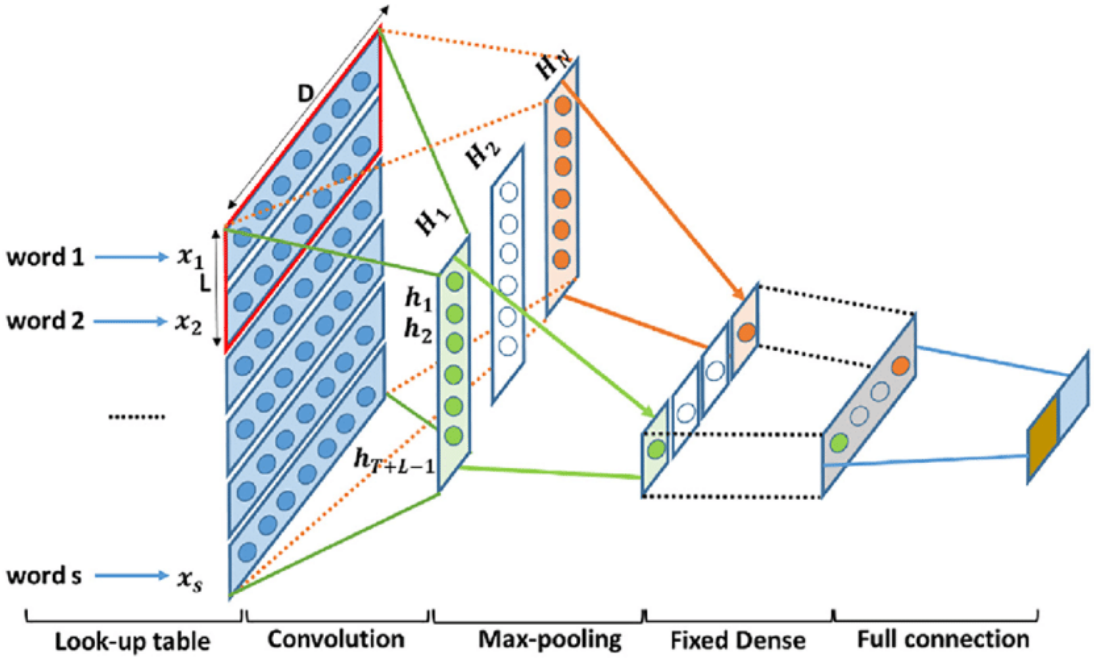

Figure 4.2. CNN architecture for text classification. Image taken from (Nguyen et al., 2019)

- Bi-GRU

  A recurrent neural network (RNN) is an extension of the feed-forward neural model to handle sequential data with different lengths. It encodes sequential data such that each node gains information from the previous node. The memory unit is integrated within the network, but it fails to keep track of the overall context when the chain becomes long. LSTM and GRU are enhanced models of RNN models with strong long-term dependency modeling capabilities that overcome this issue by having different gating mechanisms. A Bidirectional GRU (Bi-GRU) is a bidirectional RNN, that consists of a forward GRU, a backward GRU and forget gates. Figure 4.3 presents a schematic diagram of a Bi-GRU Model. The output h_t is formed based on the current input x_t and previous state h_{t-1} under the control of the two gates that determine what information is omitted or maintained to the next timestamp.

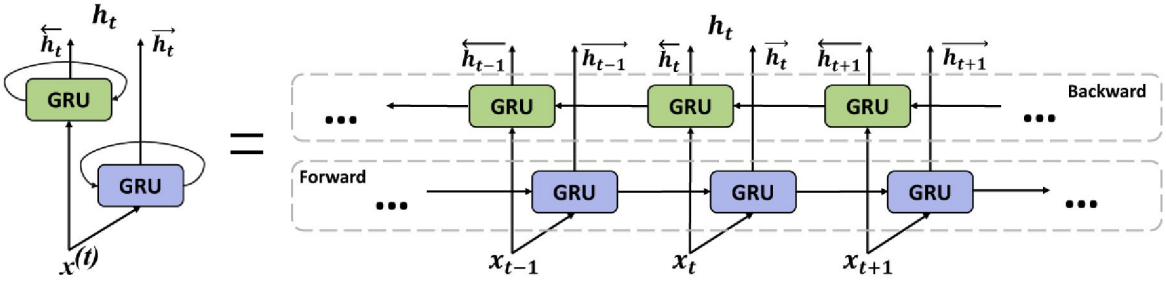

Figure 4.3. Bi-GRU model architecture. Image taken from (Liu et al., 2021)

- LSTM

  Another type of gated RNN is LSTM designed to approach the vanishing gradient problem (Hochreiter and Schmidhuber, 1997). There are four components to an LSTM cell: the cell state, the input gate (i_t), the forget gate (f_t), and the output gate (o_t) as presented in Figure 4.4. An LSTM cell c_t changes its status based on the current information x_t, hidden state h_{t-1} and cell state. The cell state is updated based on the input of activation from the previous timestamp c_{t-1} and the gates in the cell. The input gate decides which information should be added to the cell, then the forget gate determines the amount of information to be omitted or maintained throughout the sequence. Finally, the output gate decides if the memory cell's state will be passed to other memory cells using two activation functions: the sigmoid (σ) and the hyperbolic tangent (tanh) functions.

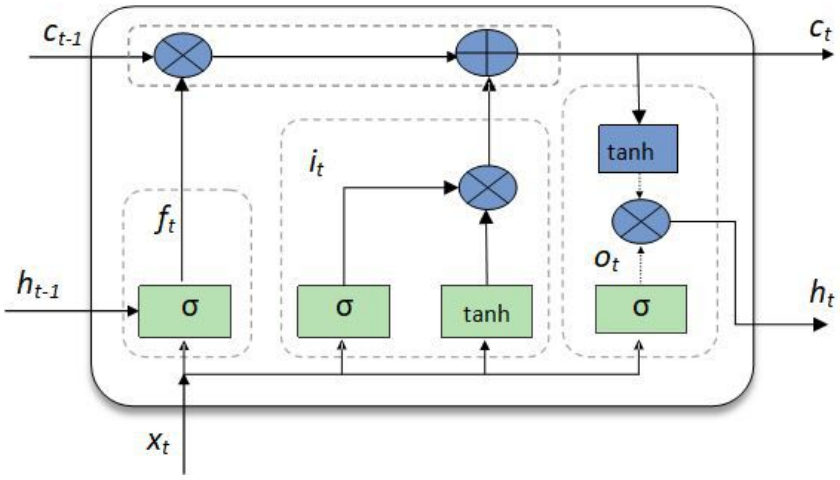

Figure 4.4. LSTM cell components

- BERT

  Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained language representation model. A stack of 12 transformers-encoder blocks with 12 multi-headed self-attention layers is the basic implementation of the BERT model. A transformer is based on a self-attention mechanism as illustrated in Figure 4.5.

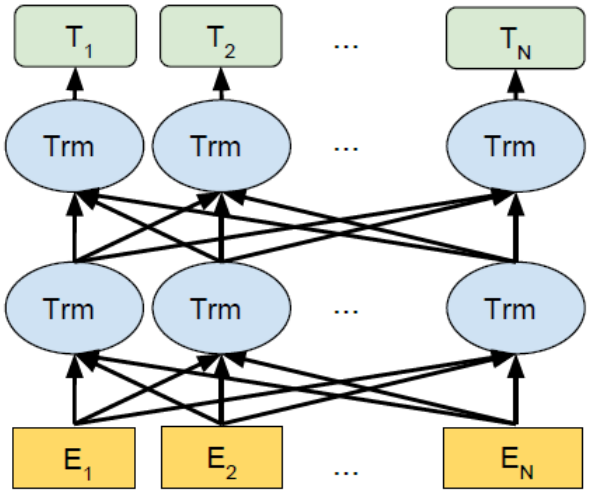

Figure 4.5. BERT bidirectional transformers layers (Devlin et al., 2018)

4.3 Results and Discussion

In this section, we evaluate five traditional machine learning algorithms and another five deep learning models for detecting depressed users based on their tweets. Accuracy, precision, recall and the F1-score are calculated using 5-fold and 10-fold cross-validation with stratified sampling to evaluate the performance of each deep learning model and traditional classifier respectively.

4.3.1 The Results of the Traditional Classifiers

This section compares SVM, LR, RF, GBDT, and XGBoost (eXtreme Gradient Boosting) classification methods. $\mathcal{D}1$ dataset was randomized, and then we used 10-fold cross-validation for training and testing as shown in Figure 4.6.

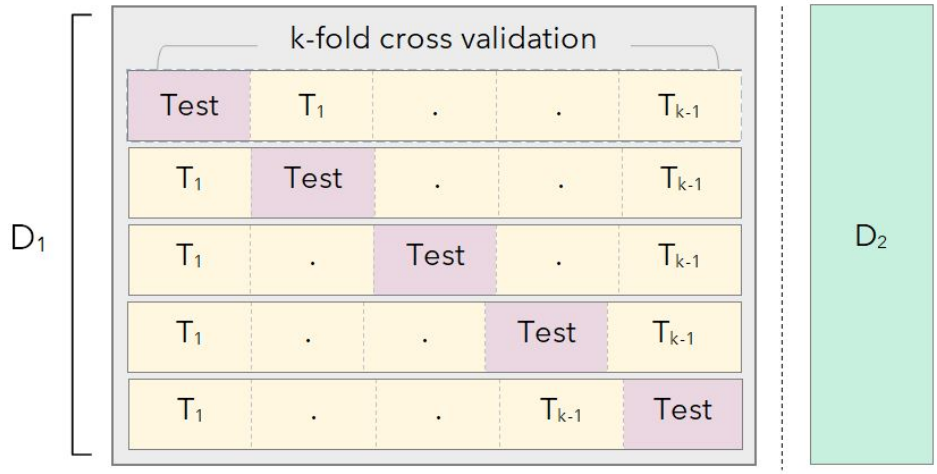

Figure 4.6. Evaluating ML models using k-fold cross-validation.
The training set D1 is split into k smaller sets. The *Test* set is used to validate the model after using the other k-1 splits for training.

We extracted the statistical features during the preprocessing phase. Then, we used 200 uni-grams with the highest tf-idf frequencies on the corpus-level weighted by IDF values to distinguish between depressed and non-depressed users[5]. We experimented with bi-grams

[5]Using TfidfTransformer from sklearn.feature_extraction package

and tri-grams as well, with no significant improvement. Besides, we used NLTK POS-tagger[6] to count the number of different types of part of speech tags for each document (user). Table 4.7 lists all the POS tags included in our POS-feature set[7]. In addition, we analyzed the output of the LIWC tool. Table 4.3 shows the highest 25 statically significant language usage differences between depressed and non-depressed users in the $\mathcal{D}1$ dataset based on Welch's t-test result. As emphasized by previous studies, we can see that the 1st person singular pronoun and present tense are significantly different between the two groups. Also the exploit of health terms (e.g., *clinic*, *diagnose*), cognitive (e.g., *know*, *think*), and certainty terms (e.g., *always*, *never*) are divergent in both groups; likewise, the usage of informal language and dictionary words is more apparent in non-depressed users than in depressed users.

CC	Coordinating conjunction	NNP	Noun, plural	TO	to	'
CD	Cardinal number	NNPS	Proper noun, singular	UH	Interjection	"
DT	Determiner	NNS	Proper noun, plural	VB	Verb, base form	#
EX	Existential there	PDT	Predeterminer	VBD	Verb, past tense	$
FW	Foreign word	POS	Possessive ending	VBG	Verb, gerund or present participle	(
IN	Preposition or subordinating conjunction	PRP	Personal pronoun	VBN	Verb, past participle	)
JJ	Adjective	PRP$	Possessive pronoun	VBP	Verb, non-3rd person singular present	.
JJR	Adjective, comparative	RB	Adverb	VBZ	Verb, 3rd person singular present	:
JJS	Adjective, superlative	RBR	Adverb, comparative	WDT	Wh-determiner	``
LS	List item marker	RBS	Adverb, superlative	WP	Wh-pronoun	
MD	Modal	RP	Particle	WP$	Possessive wh-pronoun	
NN	Noun, singular or mass	SYM	Symbol	WRB	Wh-adverb	

Figure 4.7. NLTK POS-Tags from nltk.help.upenn_tagset()

Moreover, from the Gensim[8] library, we used the LDA topic modeling tool along with the Mallet's implementation to discover L hidden topics based on coherence test. The resultant L-topic distributions are fed into the classifiers as L features. The highest coherence value was 0.46 when L is set to 50 topics. Using Scattertext and Empath packages (Fast et al., 2016), Figures 4.8 and 4.9 show some of the LDA topics in both $\mathcal{D}1$ and $\mathcal{D}2$ datasets. The x-axis represents the topic frequencies for the control users, whereas the y-axis represents the topic frequencies for the depressed users. The figures show that the *"Friends"* topic is highly associated with both groups in the two datasets. The *"Health"* and *"neglect"* topics are some of the most frequent topics used by depressed users. On the other hand, the *"fun"* and *"music"* topics are more used by the control users. Examples of

[6]http://www.nltk.org/book/ch05.html
[7]https://pythonspot.com/nltk-speech-tagging/
[8]https://radimrehurek.com/gensim

Table 4.3. Top 25 significant language differences between depressed and non-depressed (control) users based on LIWC categories ($p < 0.05$)

LIWC Category	**Control_Mean**	**Control_Std**	**Dep_mean**	**Dep_std**
WC	13160	17548	3944	5876
health	0.70	0.58	1.71	1.76
insight	1.68	0.94	2.55	1.41
cogproc	8.22	3.24	10.39	3.20
OtherP	14.43	6.13	10.96	5.39
Authentic	24.24	18.84	36.02	22.80
netspeak	5.40	2.38	4.19	2.13
Colon	5.91	4.75	3.65	4.04
focuspresent	9.36	3.20	11.00	3.24
auxverb	5.37	2.14	6.44	2.09
Dic	68.23	10.40	72.90	8.30
function	27.46	6.75	30.67	6.27
verb	13.92	3.88	15.82	3.89
AllPunc	34.13	11.83	28.75	10.47
conj	2.32	1.02	2.80	1.04
informal	7.02	2.90	5.82	2.60
Analytic	69.88	15.79	62.39	18.05
anx	0.26	0.25	0.52	0.77
i	3.45	1.83	4.33	2.20
pronoun	8.94	3.04	10.23	3.23
gender	0.47	0.50	0.67	0.47
certain	1.34	0.81	1.66	0.88
ppron	6.24	2.28	7.16	2.59
negemo	3.24	1.74	3.99	2.23
adverb	3.30	1.48	3.85	1.51

topics are illustrated in Table 4.4.

Table 4.4. Topics example for $\mathcal{D}1$ dataset

Topics	**Terms**
Health	dementia, migraine, transplant, cancer, cure, suffer, prescribed, psychologist, diarrhea, psychiatric.
Neglect	sorrow, separation, insecurity, denial, lunacy, intolerable, worsen, carelessness, sufferer, betrayal.
Musical	actor, classical, upbeat, flute, reggae, artist, collaboration, band, concert, theatre.
Fun	arcade, trampoline, celebration, silly, themed, prank, excite, favorite, laugh, play.

The evaluation results using traditional ML techniques where each feature set is confined are illustrated in Table 4.5. LIWC features have the best results in general, then LDA topics, Tf-idf, and POS features. The statistical features did not yield good results.

In the next set, we present the experiments of word embeddings applying different averaging techniques. We used the Gensim library which contains even more options for training word vectors than just Word2vec, such as fastText and Doc2Vec. We applied the following strategies to represent the users' posts as features employing word embeddings. The simplest method is through "Averaging on Word Embeddings" achieved by calculating the unweighted average of the 200-dimensional embeddings using Word2vec of all the words that appeared at least three times in the text, and similarly "Tf-idf average on Word Embeddings" is calculated using the mean of the idf weight of the Tf-idf model. The other method uses Doc2Vec model that utilizes Word2vec model and attaches a paragraph-ID vector to each paragraph or document. We used the (PV-DM) Paragraph Vector with Distributed Memory model (Le and Mikolov, 2014) with five negative sampling and 200-dimensions. The paragraph-ID and word vectors weights are trained in an unsupervised manner, allowing the paragraph-ID vector to hold a vector representing the document's context.

The fourth feature set uses 200-dim GloVe pre-trained vectors on 2B tweets. The results were comparable using the four generated features based on word embeddings, but

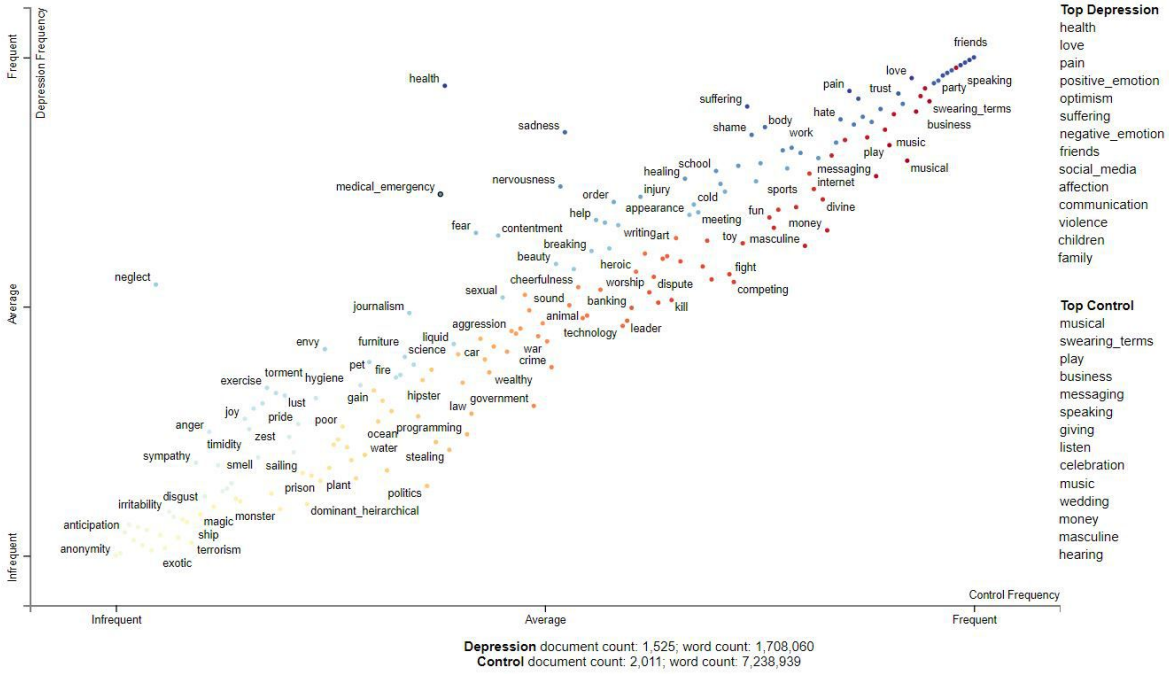

Figure 4.8. Topic visualization for terms used by depression and control users within $\mathcal{D}1$

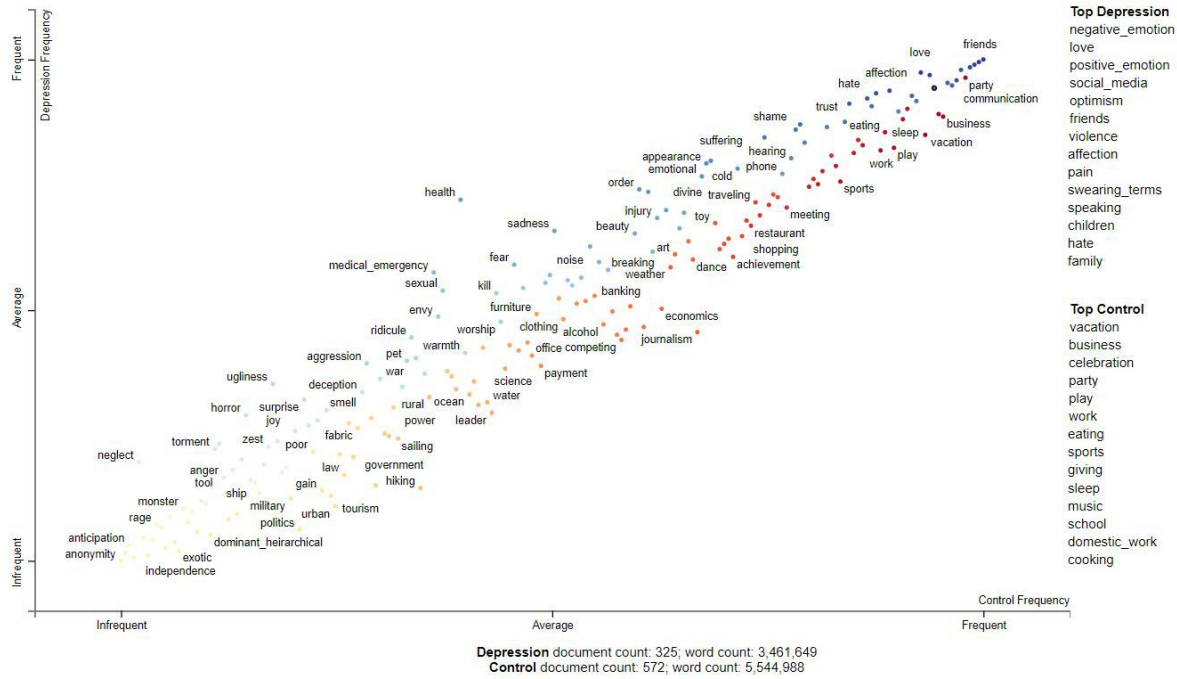

Figure 4.9. Topic visualization for terms used by Depression and Control users within $\mathcal{D}2$

Table 4.5. Comparison of various ML algorithms using engineered features on $\mathcal{D}1$ dataset (10 cross-validation)

Features	**Model**	**Acc.%**		**Pre.**	**Rec.**		**F1**
LIWC	SVM	80.1	(0.67-0.92)	0.773	0.761	(0.59-0.92)	0.767
	LR	80.0	(0.67-0.92)	0.774	0.762	(0.60-0.91)	0.767
	RF	81.3	(0.69-0.91)	0.785	0.777	(0.62-0.91)	0.781
	GBDT	81.9	(0.69-0.94)	0.796	**0.781**	(0.60-0.93)	**0.788**
	XGBoost	**82.1**	(0.69-0.95)	**0.801**	0.776	(0.61-0.95)	**0.788**
POS	SVM	76.1	(0.64-0.92)	0.717	0.762	(0.59-0.92)	0.736
	LR	76.2	(0.64-0.92)	0.719	0.760	(0.60-0.91)	0.736
	RF	78.7	(0.69-0.91)	0.750	0.762	(0.62-0.91)	0.755
	GBDT	80.1	(0.69-0.94)	0.774	0.760	(0.60-0.93)	0.766
	XGBoost	79.6	(0.69-0.95)	0.769	0.755	(0.61-0.95)	0.761
Tf-idf	SVM	77.4	(0.64-0.92)	0.735	0.767	(0.59-0.92)	0.748
	LR	77.6	(0.64-0.92)	0.738	0.765	(0.60-0.91)	0.749
	RF	80.0	(0.69-0.91)	0.770	0.770	(0.62-0.91)	0.769
	GBDT	80.8	(0.69-0.94)	0.782	0.771	(0.60-0.93)	0.776
	XGBoost	80.8	(0.69-0.95)	0.782	0.769	(0.61-0.95)	0.775
Topics	SVM	78.2	(0.64-0.92)	0.749	0.761	(0.59-0.92)	0.752
	LR	78.4	(0.64-0.92)	0.752	0.762	(0.60-0.91)	0.755
	RF	80.2	(0.69-0.91)	0.776	0.767	(0.62-0.91)	0.770
	GBDT	81.6	(0.69-0.94)	0.794	0.775	(0.60-0.93)	0.784
	XGBoost	81.5	(0.69-0.95)	0.792	0.775	(0.61-0.95)	0.783
Sents & Emo.	SVM	75.6	(0.60-0.92)	0.729	0.690	(0.26-0.92)	0.700
	LR	75.8	(0.59-0.92)	0.730	0.694	(0.28-0.91)	0.704
	RF	78.0	(0.63-0.91)	0.753	0.728	(0.45-0.91)	0.739
	GBDT	79.1	(0.63-0.94)	0.765	0.744	(0.56-0.93)	0.754
	XGBoost	79.3	(0.65-0.95)	0.768	0.743	(0.56-0.95)	0.755

the recall did not exceed 0.77. The results using various word embedding aggregation strategies are demonstrated in Table 4.6.

Table 4.6. Comparison of various ML algorithms using word embedding features

Features	**Model**	**Acc.**		**Pre.**	**Rec.**		**F1**
Averaging on	SVM	77.6	(0.60-0.92)	0.750	0.718	(0.26-0.92)	0.725
Word	LR	77.7	(0.59-0.92)	0.751	0.720	(0.28-0.91)	0.729
Embeddings	RF	78.8	(0.63-0.91)	0.766	0.733	(0.45-0.91)	0.748
	GBDT	80.0	(0.63-0.94)	0.778	0.753	(0.56-0.93)	0.765
	XGBoost	80.5	(0.65-0.95)	0.783	0.755	(0.56-0.95)	0.768
Tf-idf Avg.	SVM	80.8	(0.60-0.95)	0.786	0.760	(0.26-0.93)	0.767
on Word	LR	80.9	(0.59-0.94)	0.788	0.762	(0.28-0.93)	0.769
Embeddings	RF	79.1	(0.63-0.91)	0.775	0.728	(0.45-0.91)	0.749
	GBDT	81.0	(0.63-0.94)	0.792	0.761	(0.56-0.93)	0.775
	XGBoost	**81.7**	(0.65-0.95)	**0.802**	**0.765**	(0.56-0.95)	**0.782**
DocVec	SVM	79.2	(0.60-0.95)	0.769	0.738	(0.26-0.93)	0.746
	LR	79.3	(0.59-0.94)	0.770	0.740	(0.28-0.93)	0.749
	RF	78.7	(0.63-0.91)	0.767	0.727	(0.45-0.91)	0.745
	GBDT	80.3	(0.63-0.94)	0.780	0.756	(0.56-0.93)	0.767
	XGBoost	80.8	(0.65-0.95)	0.788	0.757	(0.56-0.95)	0.772
Pre-trained	SVM	80.2	(0.60-0.95)	0.780	0.750	(0.26-0.93)	0.758
GloVe	LR	80.3	(0.59-0.94)	0.781	0.752	(0.28-0.93)	0.761
	RF	78.8	(0.63-0.91)	0.772	0.724	(0.45-0.91)	0.745
	GBDT	80.6	(0.63-0.94)	0.785	0.757	(0.56-0.93)	0.770
	XGBoost	81.2	(0.65-0.95)	0.796	0.759	(0.56-0.95)	0.776

ML algorithms were tuned using different parameters including the depth of the tree, learning rate and the loss function using grid search to find the optimal parameters for different combinations of features. Other methods can be used for fine tuning such as random search and Bayesian optimization. The hyperparameters of the GBDT model were tuned to the following tree-specific hyperparameters: *min_samples_leaf* to 20 samples, *max_depth* of each regression estimator to 15 splits and *max_features* to 50. The boosting-related hyperparameters include: *n_estimators* of 79 sequential trees, *subsample* of 0.8 and

learning_rate of 0.05. Table 4.7 part(A) shows the best results on $\mathcal{D}1$ dataset gained from combining different engineered features while Table 4.7 part(B) shows the evaluation on $\mathcal{D}2$ dataset.

Finally, $\mathcal{D}2$ dataset was used to test all the models that were trained on $\mathcal{D}1$ dataset. Table 4.7 Part (B) shows the results of the best performing models. Although XGBoost has the best recall of 0.874, GBDT has the best F1-score of 0.882 with the second best recall score of 0.870. In addition, GBDT has the highest precision, recall and F1-score on $\mathcal{D}2$ using only half of the features.

The test results show that the models generalize well on unseen dataset and can be applied on similar population representative dataset. As mentioned in Section 3.4, $\mathcal{D}1$ and $\mathcal{D}2$ datasets are different in collection, annotation and anonymization process. $\mathcal{D}2$ was verified by a human annotator to remove misleading statements such as jokes or citations whereas the training dataset $\mathcal{D}1$ depends only on users self declaration. The number of tweets for each user and the length of the tweets differ between the two datasets as well. Figure 4.10 shows the number of posts distributions for both datasets. For dataset $\mathcal{D}1$ the median number of tweets for the 897 users is 3000 and the mean is 2222, whereas for $\mathcal{D}2$ the median number of tweets for the 3536 users is 162 and the mean is 437.

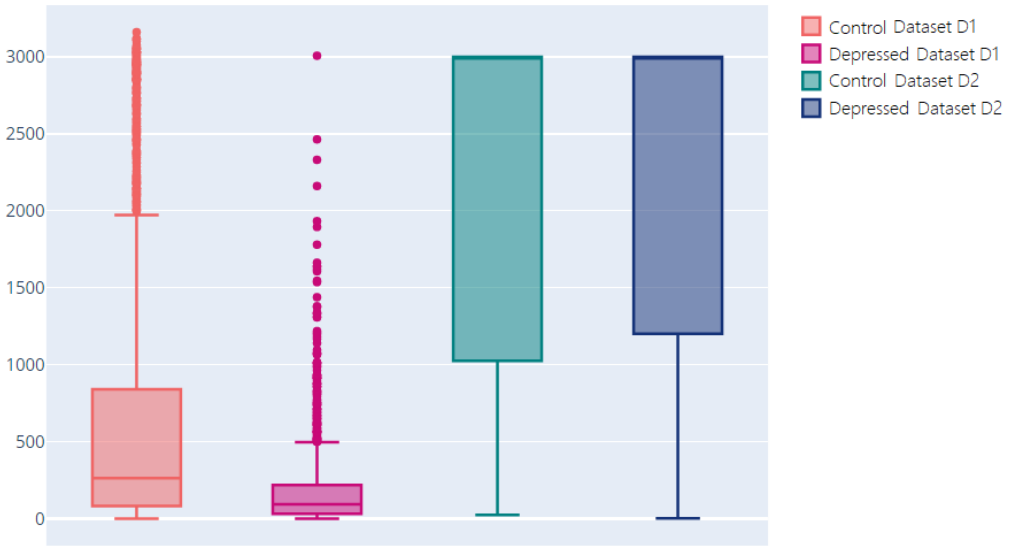

Figure 4.10. Boxplot of number of tweets for $\mathcal{D}1$ and $\mathcal{D}2$ datasets

In addition, the term frequency in both datasets differs. Figure 4.11 shows the top 10 used terms in the two datasets for each user category. In the $\mathcal{D}1$ dataset, we notice

Table 4.7. Comparison of various ML algorithms using different features combinations

Part(A) Combined features results on $\mathcal{D}1$ dataset using 10-fold cross-validation						
Features	**# of feat.**	**Model**	**Acc.%**	**Pre.**	**Rec.**	**F1**
LIWC+	293	SVM	88.6	0.863	0.875	0.869
Doc Tf-idf		LR	88.7	0.869	0.870	0.869
		RF	83.7	0.828	0.788	0.807
		GBDT	87.4	0.875	0.826	0.849
		XGBoost	88.1	0.875	0.847	0.860
LIWC+POS+	588	SVM	93.1	0.913	0.928	0.920
Tfidf+Topic+		LR	93.9	0.921	0.940	0.930
Tf-idf_Mean_WE		RF	93.5	0.930	0.919	0.924
		GBDT	96.0	0.956	0.951	0.953
		XGBoost	96.4	**0.960**	0.958	0.958
All features	1211	SVM	95.1	0.938	0.950	0.944
		LR	95.9	0.949	0.956	0.952
		RF	92.7	0.923	0.906	0.914
		GBDT	95.9	0.959	0.945	0.952
		XGBoost	**96.5**	**0.960**	**0.961**	**0.960**
Part (B) Combined features results on $\mathcal{D}2$ dataset						
LIWC + POS +	588	SVM	90.4	0.876	0.844	0.837
Tfidf + Topic +		LR	91.2	0.883	0.855	0.846
Tf-idf_Mean_WE		RF	90.8	0.882	0.837	0.844
		GBDT	92.9	**0.904**	**0.872**	**0.879**
		XGBoost	**93.4**	0.872	0.871	0.871
All features	1211	SVM	92.0	0.882	0.870	0.869
		LR	92.6	0.891	0.877	0.877
		RF	90.1	0.894	0.826	0.836
		GBDT	92.8	**0.912**	0.870	**0.882**
		XGBoost	**93.5**	0.873	**0.874**	0.873

that there are more frequent use of the pronouns: *i, you, we, they* and *me* by the control and the depressed groups. But only the first user pronoun *i* is the the most used in the $\mathcal{D}2$ dataset. In the same time, in $\mathcal{D}1$, the average usage of the first user pronoun is 197 per depressed user versus 170 on average for non-depressed user, confirming trends from previous research in this field (Mowery et al., 2017b; Lyons et al., 2018; Wu et al., 2020).

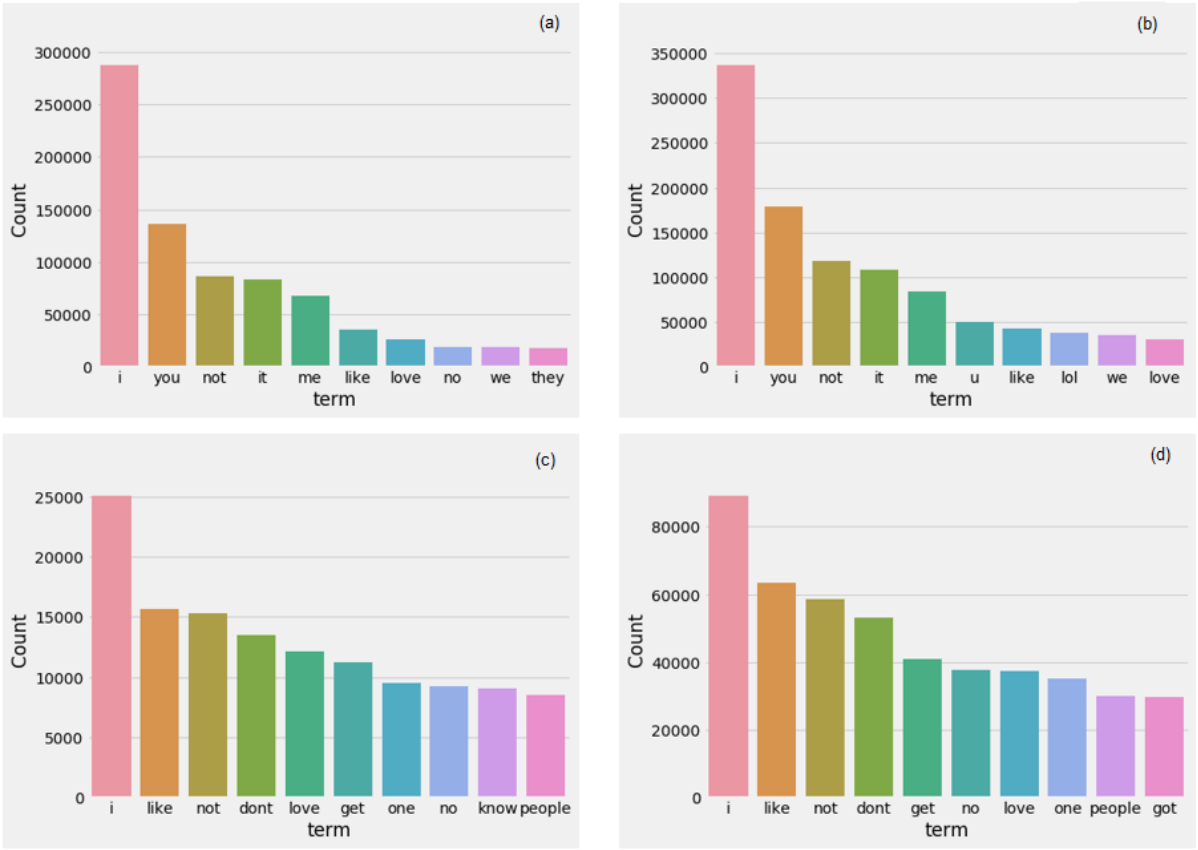

Figure 4.11. Differences in term frequency between $\mathcal{D}1$ (a: Depressed) & (b: Control) and $\mathcal{D}2$ (c: Depressed) & (d: Control) datasets

4.3.2 Deep Learning Models

Our baseline is Neural-Net Language Model (NNLM) based on a neural probabilistic language model (Bengio et al., 2003). It takes the user's tweets as input represented by a matrix $258,732 \times 32$ ($max\_words\_in\_Users'\_post \times batch\_size$). Then, the embedding layer starts with random weights, and the weights are modified during the training phase like any other layer to minimize the loss function (binary_crossentropy) using the (adam) optimization method. We reached an accuracy of 80.4% and an F1-score of 0.759 after 100 epochs.

Then, as a part of transfer learning, we used a 128-dimension token-based text embedding trained on the English Google News 200B corpus sentence embedding[9] to capture the semantic content of the post instead of the randomly initialized embedding. NNLM is trained to learn vector representations of words concurrently and predict the next word given the preceding words' representations. The sentence embedding is generated by adding up all the word vectors then dividing by the square root of the sum of the word-vectors' squares.

Afterwards, we used Bi-GRU and Bi-LSTM models with training word embedding or using pre-trained word embeddings. After tokenizing users' posts, the words that represent each user were placed in an indexed dictionary, then a sequence of indices for the words were fed to the embedding layer in an LSTM/GRU network using pre-trained word embeddings (GloVe[10], fastText[11] trained on Wikipedia pages, fastText using 300-dim word embedding of 2 million word vectors trained on Common Crawl (600B tokens) and NumberBatch[12]). Based on the recall and F1-score results on the test dataset ($\mathcal{D}2$), the best generalizable model is CNN using fastText word embeddings with a recall of 0.914 and F1-score of 0.898.

We trained a CNN model on $\mathcal{D}1$ dataset, inspired by (Kim, 2014). The 1D CNN model is trained on a top of pre-trained word embeddings on sentence level using 300 dimension fastText and GloVe-Twitter trained on 2B tweets and containing a total of 27B uncased tokens and 1.2M vocab with different dimensions: 50d, 100d, & 200d vectors.

[9] https://tfhub.dev/google/nnlm-en-dim128/2
[10] https://nlp.stanford.edu/projects/glove/
[11] https://fasttext.cc/
[12] https://github.com/commonsense/conceptnet-numberbatch

To capture the most important features we apply the max-overtime pooling operation. The architecture of the model is presented in Figure 4.12. Different hyper parameters were applied to tune the model including the number of filters, different filter sizes and max number of features. We experimented with different embedding dimensions (50,100,200,300), filter sizes (3,5,7), number of filters (32,64) and dropout probabilities (0.1-0.5). Using 5-fold cross-validation fastText-Crawl model achieved an F1-score of 0.84 and a recall of 0.89 but when tested on the $\mathcal{D}2$ dataset, F1-score increased to 0.90 and the recall of 0.91. This shows that the model is generalizable and can be applied on the $\mathcal{P}1$ dataset to infer the depression within the population level.

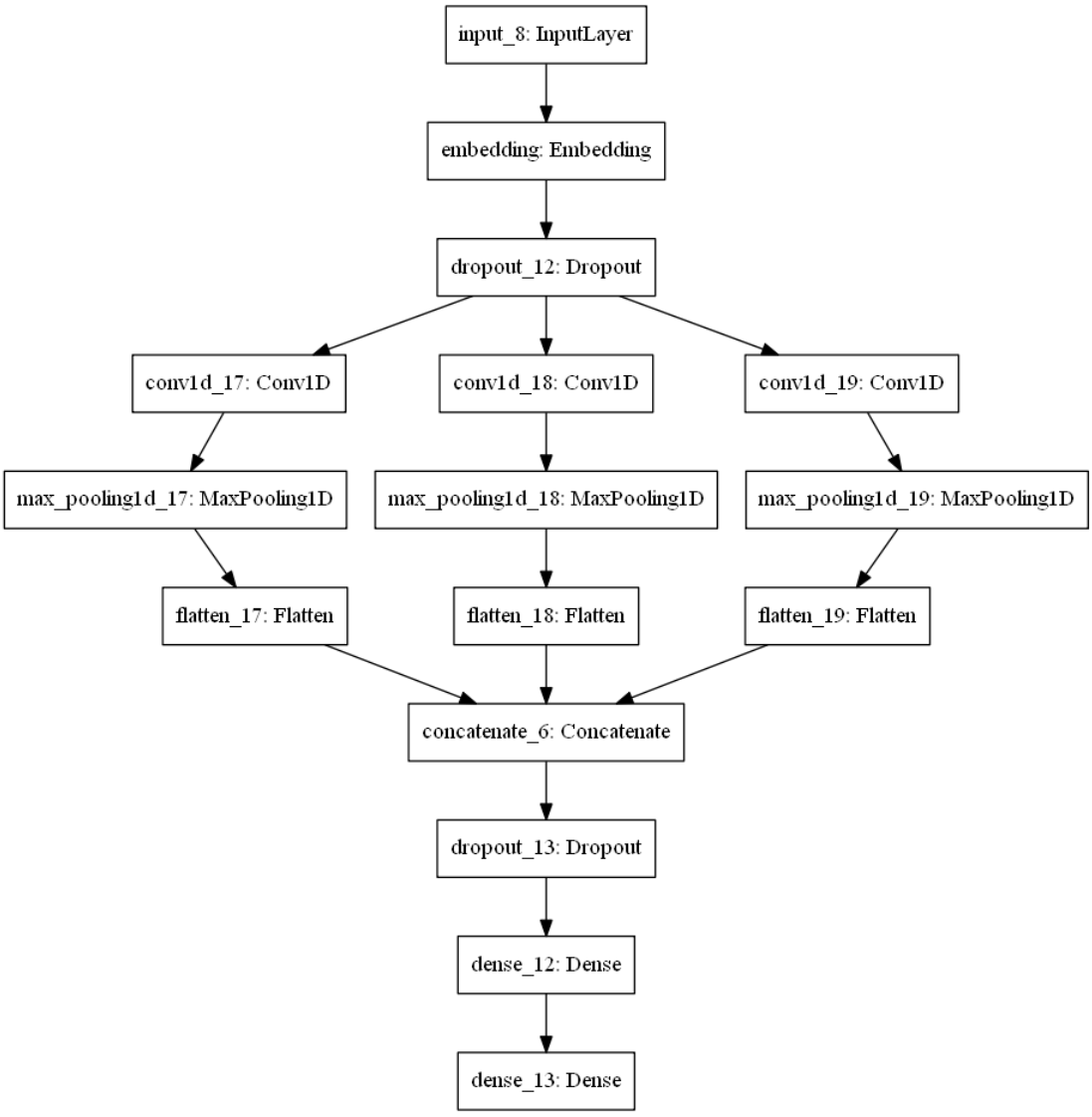

Figure 4.12. 1-D CNN architecture, with three convolution layers of filter sizes 3,5,7

The BERT model as shown in Figure 4.13 did not perform well as the text is too long, and the training was based on 512 tokens only. The median length of the tokens that represent users in the $\mathcal{D}1$ dataset is 13,764 tokens, and the maximum length is 113,214 tokens. In the training phase, 95.4% of the training data was not included because it exceeded the limit. For BERT to generalize well, another method needs to be integrated with the BERT model, such as dividing the user's text $\mathcal{T}$ into *(512)* chunks, then aggregating the output of each chunk or applying a voting mechanism. Also, striding through the tokens of the users' posts should prevent each chunk from being treated as an independent text.

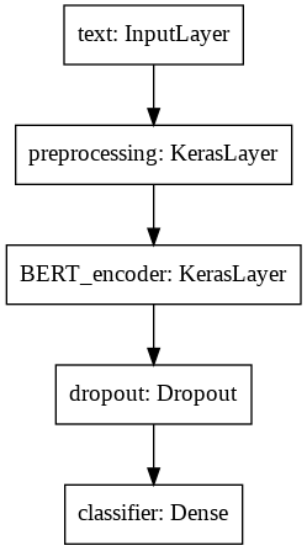

Figure 4.13. BERT-Model using Tensorflow Hub (Abadi et al., 2015)

Table 4.8 Part (A) and (B) summarizes the results of these experiments. Part (A) shows the output of the 5-fold stratified cross-validation to make sure that each fold can represent the data on the $\mathcal{D}1$ dataset. Then, the models were tested on the $\mathcal{D}2$ dataset as shown in Table 4.8 Part (B). The results show that although GBDT was the best generalizable model among the traditional machine learning methods using half of the training features, as illustrated earlier in Table 4.7 with 0.872 recall and 0.879 F1-score, the CNN deep learning model with fastText-Crawl word embeddings (CNN fastText-Crawl) exceeds GBDT in terms of recall(0.914) and F1-score(0.898) showing that CNN fastText-Crawl maintains its good performance on D2 and can be used to predict depression within the Canadian population, as will be explained in the next section.

Table 4.8. Comparison of various deep learning algorithms

Part (A) On $\mathcal{D}1$ dataset (5-fold cross-validation)					
Embeddings	**Model**	**Accuracy%**	**Precision**	**Recall**	**F1-score**
Related work (Orabi et al., 2018)		88.0	0.874	0.870	0.870
Word2vec	Bi-GRU	74.2	0.698	0.543	0.577
NumberBatch	Bi-GRU	71.7	0.762	0.394	0.469
GloVe	Bi-GRU	75.0	0.705	0.576	0.618
	CNN	**85.1**	**0.793**	0.423	0.538
fastText-Wiki	Bi-GRU	74.8	0.713	0.490	0.564
fastText-Crawl	Bi-GRU	73.9	0.721	0.486	0.556
	LSTM	76.1	0.719	0.628	0.630
	CNN	**85.1**	**0.793**	**0.885**	**0.836**
BERT-uncased	BERT	78.4	0.771	0.675	0.737
Part (B) Test on Dataset $\mathcal{D}2$					
Word2vec	Bi-GRU	83.5	0.873	0.724	0.791
NumberBatch	Bi-GRU	78.8	0.874	0.593	0.707
GloVe	Bi-GRU	83.1	0.736	**0.949**	0.829
	CNN	88.8	0.823	0.942	0.879
fastText-Wiki	Bi-GRU	82.7	0.793	0.812	0.802
fastText-Crawl	Bi-GRU	89.5	0.837	0.914	0.886
	LSTM	76.0	0.755	0.657	0.703
	CNN	**91.0**	**0.882**	0.914	**0.898**
BERT-uncased	BERT	67.0	0.559	0.452	0.500

4.4 Depression Detection at Population Level

To examine the association between our model's prediction and official statics, we used the 2015-2016 Canadian Community Health Survey (CCHS). The CCHS is a cross-sectional survey that provides information relating to health at federal and provincial levels conducted by Statistics Canada. The depression module was optional content in the 2015; therefore only seven provinces and one territory participated *(the number of respondents was n = 52,996)* as shown in Table 4.9. Depressive symptoms were assessed using the Patient Health Questionnaire-9 (PHQ-9). Considering the depression scale for PHQ-9 > 9, the 12-month prevalence rate for depression in Canada for 2015 was estimated to be 7.1%. The number of users included in our sample *(p = 24,251)* is limited to the provinces included in the survey. We applied the CNN model "fastText-Crawl" on our p sample, as it achieved the best accuracy, precision and F1-score on $\mathcal{D}2$ among the other models with 91%, 0.882 and 0.898. The recall of the model was the best on the training dataset $\mathcal{D}1$.

Table 4.9. Geographic and sex population difference between CCHS 2015-2016 data and $\mathcal{P}1$

Pr.	$\mathcal{P}1_M$	$\mathcal{P}1_F$	n_M	n_F	*Diff.%*	D_M	D_F	PD_M	PD_F
NL	324	366	1,396	1,660	1.54	7.8	13	7.4	12.8
PE	223	231	720	1,013	1.02	7.4	8.3	7.3	9.0
NS	1,015	1,060	2,031	2,473	1.32	13.8	15.7	9.6	13.7
NB	435	370	1,354	1,746	6.34	12.5	14.8	9.4	13.5
ON	9,257	7,893	13,933	16,370	5.08	8.7	12.4	9.2	12.6
MN	806	687	2,327	2,708	4.79	8.6	13	8.6	13.1
SK	819	574	2,020	2,300	8.29	8.7	12.2	9.3	12.2
NT	101	90	477	468	2.48	11.3	7.9	9.3	11.1

$\mathcal{P}1_{M/F}$ the number of male/female users in P1 Dataset,

$n_{M/F}$ the number of sample users in CCHS 2015-2016,

$D_{M/F}$ the estimated proportion with mild to severe depression based on CCHS 2015-2016 (PHQ-9)

$PD_{M/F}$ the predicted proportion of male/female users based on fastText-Crawl-CNN model

The association between the prediction output (PD) and the CCHS data is calculated using Pearson correlation coefficient (ρ) based on the following equation:

$$Correl(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Table 4.10 shows the calculated Pearon correlation coefficient to measure of the strength of the relationship between PD and CCHS results on sex, age, and province levels. The Pearson correlation coefficients that measure the strength of the relationship between the CCHS results and the depression prediction using CNN-fastText-Crawl model on $\mathcal{P}1$ dataset range between 0.50 and 0.99.

Table 4.10. Pearson linear correlation between CCHS 2015-2016 and the prediction on $\mathcal{P}1$ dataset dataset based on CNN-fastText-Crawl model

	Pearson Correlation
Sex	
Female	0.93
Male	0.95
Age	
<25	0.95
25-34	0.95
35-44	0.93
45-54	0.96
55-64	0.99
≥65	0.97
Province	
Manitoba	0.71
New Brunswick	0.50
Newfoundland and Labrador	0.66
Northwest Territories	0.69
Nova Scotia	0.86
Ontario	0.62
Prince Edward Island	0.57
Saskatchewan	0.91

Figures 4.14 and 4.15 show the distribution of predicted depressed users in the $p \in \mathcal{P}1$ dataset that was described in Section 3.4.1 and the mapping population from CCHS for females and males. It shows that the predictions at user-level correlate with the survey findings with differences $\leq 0.7\%$ except for the provinces of Nova Scotia and New Brunswick, where the prediction is lower in both sexes (more apparently in males). The depression among female respondents is higher than in male respondents, according to CCHS and PD predictions, except for the Northwest Territories, where the numbers of users are the lowest in the $\mathcal{P}1$ dataset as shown in Figure 4.16. Appendix A.4 shows the demographics of the estimated users in PD reference to the CCHS survey.

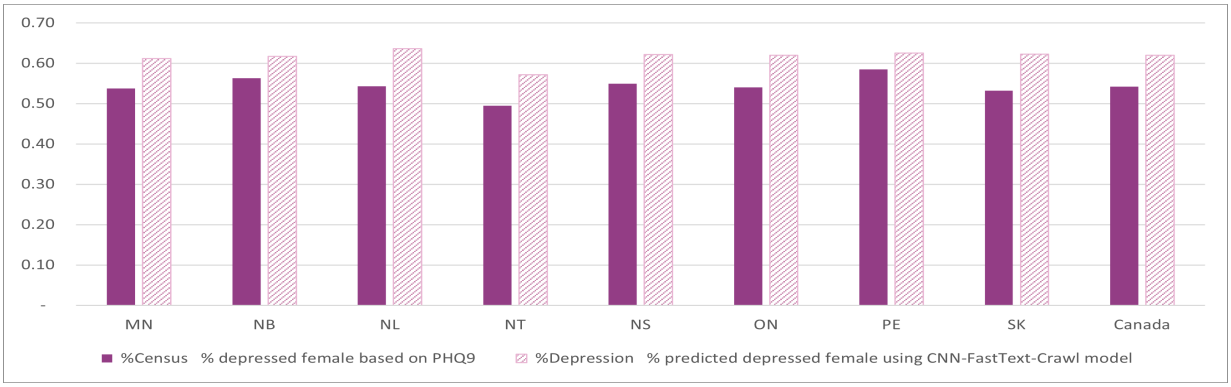

Figure 4.14. Relevance of predicted depressed female users using CNN-fastText model on $\mathcal{P}1$ dataset to CCHS 2015

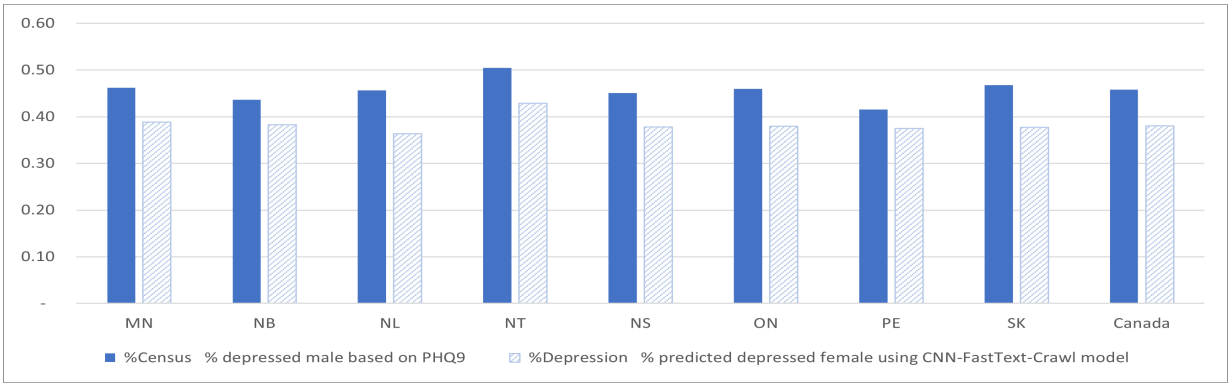

Figure 4.15. Relevance of predicted depressed male users using CNN-fastText model on $\mathcal{P}1$ datasetto CCHS 2015

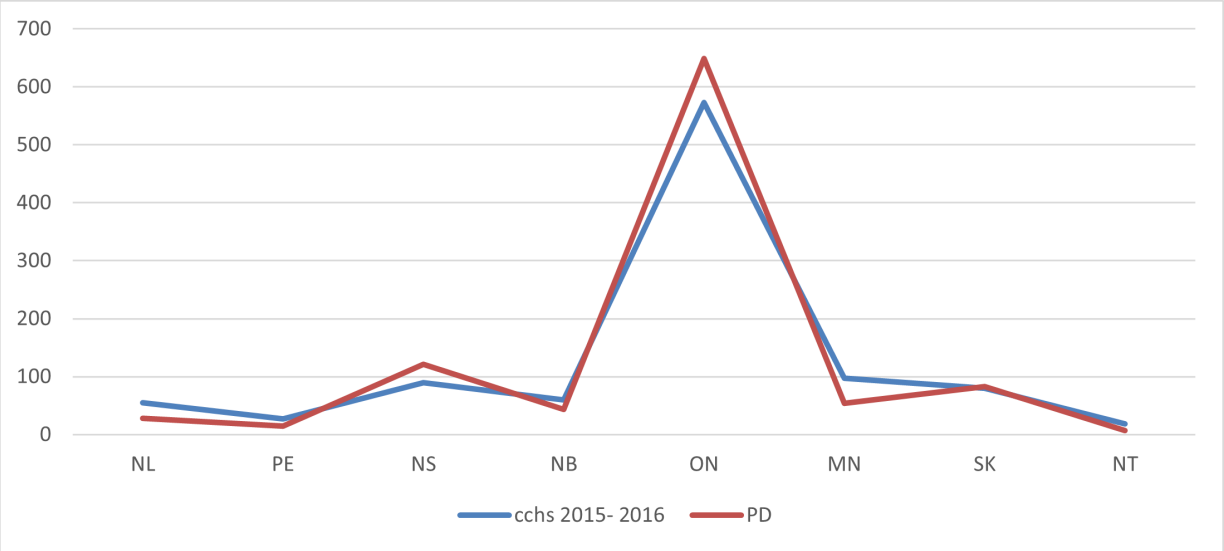

Figure 4.16. Depressed estimation based on CNN-fastText model on $\mathcal{P}1$ relevant to CCHS 2015-2016 survey

4.5 Summary

In this chapter, we experimented with two Twitter datasets to build an optimal model for population-level prediction. We discussed the experiments we performed for training user-level classifiers. After preprocessing the data, we experimented with different features and machine learning algorithms. We chose the "best" trained model to be the best generalizable model, that achieves the best recall and F1-score on both the training and the test dataset. This model was applied on the population dataset and the results were compared with Canadian statistics on depression. In the next chapter, we perform similar experiments to detect suicide ideation within the Canadian population.

Chapter 5

Suicide Ideation from User to Population

5.1 Overview

In this chapter[1], we utilize personal narratives collected through the popular social media website (Reddit) to build a model suitable for predicting suicide ideation in a sample of Twitter users that is representative for the Canadian population. The suitability of the model is measured by the F1-score and recall value on a Twitter dataset. These two measures were chosen because the datasets in our study are not balanced. We want high F1-score so that the model correctly identifies depressed users ("True-positives"), with high recall in order to minimize the number of missed depressed users ("False-Negative"). For suicide ideation and similar critical scenarios, we need to assess the model's ability to identify positive cases (users with suicide ideation) where missing a case is more critical than wrongly classifying the case as positive (specially from the population-level perspective).

The task is defined as follows: given a set of reddit users $\{U_1, U_2, ..U_n\}$ such that each user posted a variant number of posts, all the posts of the user is represented by x_i features vector and have one binary label y_i. We want to learn a classification function $f : x \rightarrow y$ that maps the user to his/her label; i.e. $y = \{'at\text{-}risk', 'not\text{-}at\text{-}risk'\}$.

[1]Parts of this chapter appeared as Skaik and Inkpen (2020c)

In Section 5.2 we list the feature sets used in our models followed by the results obtained using traditional and deep learning classification methods in Section 5.3. After the discussion of the results, we apply the chosen prediction model to the population level and compare it with CCHS survey in Section 5.4.

5.2 Suicide Ideation Classification Models

In this chapter, we include the evaluation of four traditional machine learning models and four deep learning models for identification of suicidal users based on their posts/tweets. Other models were tried, but their results are not shown due to poor performance.

Accuracy, precision, recall and the F1-score are calculated to assess the performance of each classifier using 5-fold cross-validation with stratified sampling.

5.2.1 Feature Selection

Before extracting the features from the text, we performed several pre-processing steps to reduce the noise from the original data as mentioned in Section 3.4.4. We performed the following on the $\mathcal{S}1$ Reddit dataset: we removed the special deidentification tokens including PERSON and URL (as the data was anonymized to preserve the privacy of users by automatically replacing any mentions of specific identity related information such as email addresses, URLs, and names with unified related tokens). We concatenated the title and body of the text. For all datasets we performed the following: lowercased the text, removed the punctuation and tokenized them using SpaCy[2], we extracted the emoticons and emojis from the text, removed non-alphanumeric characters, performed stopwords removal (except pronouns and negation) then ordered the posts in chronological order. Then, for all the datasets, we concatenated all the posts of each user into a single document that represents the user. We extracted the features from the users' posts. We experimented with the same features that were mentioned in Section 4.2 and we are reporting the most significant features based on the experiments results. Following are the selected sets of features used, some of them were motivated by (Ji et al., 2018).

[2]https://spacy.io

- **Statistical Features:** We start our experiments by extracting statistical features from each document. The statistical features include the document length (including title and body length for Reddit users), word counts, number of punctuation marks, emojis and emoticons used, frequency of posting, as well as number of posting during the different periods of the day: morning, noon, evening and night.

- **Tf-idf:** We calculated the term frequency-inverse document frequency (tf-idf) values for unigram and bigrams. Tf-idf represents the significance of a term in a document based on the frequency of its appearance in the document, and inversely proportional to its frequency in the rest of the corpus (because a term that appear in fewer documents is more useful for classification). The initial number of features was 7,637 and 130,517 for uni-grams and bi-grams, respectively. We used Principal Component Analysis (PCA) for dimensionality reduction that reduced the tf-idf features to 325 features.

- **Linguistic Features:** We used the 93 Linguistic features that are extracted using the LIWC tool - as described in Section 4.2. Table 5.1 shows the highest 25 statically significant language usage differences between users with the risk-of suicide and the non-at-risk users in $\mathcal{S}1$ dataset based on Welch's t-test result.

- **Topic Modeling:** We used LDA topic modeling tool to extract L hidden topics based on a coherence test as described in Section 4.2. The resultant L-topic distributions are fed into the classifiers as L features. In our case, the highest coherence value was 0.72 when L is set to 20 topics. Figures 5.1 and 5.2 illustrate the topics in the $\mathcal{S}1$ and $\mathcal{S}2$ datasets using the Scattertext visualization tool. The x-axis represents the topic frequencies for the control users whereas the y-axis represents the topic frequencies for the suicidal users. Figure 5.1 shows that there are slight differences between the topics discussed by control and suicidal users in $\mathcal{S}1$ dataset. However, in $\mathcal{S}2$ dataset, topics such as *war, sadness, kill, fear, nervousness ..etc* are more used by suicidal users, and topics such as *achievement, fun, celebration, play, sport* and *travelling* are more discussed by the control group.

- **Emotion Features:** We also used sentiment analysis VADER tool for sentiment analysis as described in Section 4.2.

Table 5.1. Top 25 significant language differences between users at-risk of suicide and non-at-risk (control) users based on LIWC Categories ($p < 0.05$)

LIWC Cat.	**Ctrl_mean**	**Ctrl_std**	**Suicide_mean**	**Suicide_std**
Authentic	37.59	31.77	80.22	21.83
i	7.11	3.56	11.76	2.95
negemo	2.44	1.75	3.88	1.83
WC	202.72	215.75	459.69	610.14
negate	2.26	1.35	3.13	1.56
WPS	7.40	3.72	9.77	4.75
anger	0.60	0.87	1.11	1.04
verb	17.98	4.05	19.93	3.21
swear	0.20	0.51	0.49	0.70
sad	0.74	0.84	1.17	0.97
Dic	82.98	9.52	87.29	7.63
auxverb	9.29	2.85	10.57	2.44
bio	2.00	1.83	2.81	1.56
adj	3.34	1.80	4.13	1.64
time	4.48	2.39	5.52	2.24
relativ	10.70	3.73	12.28	3.39
Apostro	3.03	2.25	3.99	2.24
function	53.03	7.14	55.80	6.15
affect	5.38	2.63	6.37	2.11
sexual	0.11	0.37	0.26	0.49
focuspresent	12.90	4.13	14.37	3.62
certain	1.48	1.28	1.93	1.11
health	1.12	1.28	1.57	1.23
compare	1.79	1.29	2.24	1.24
risk	0.57	0.70	0.78	0.68

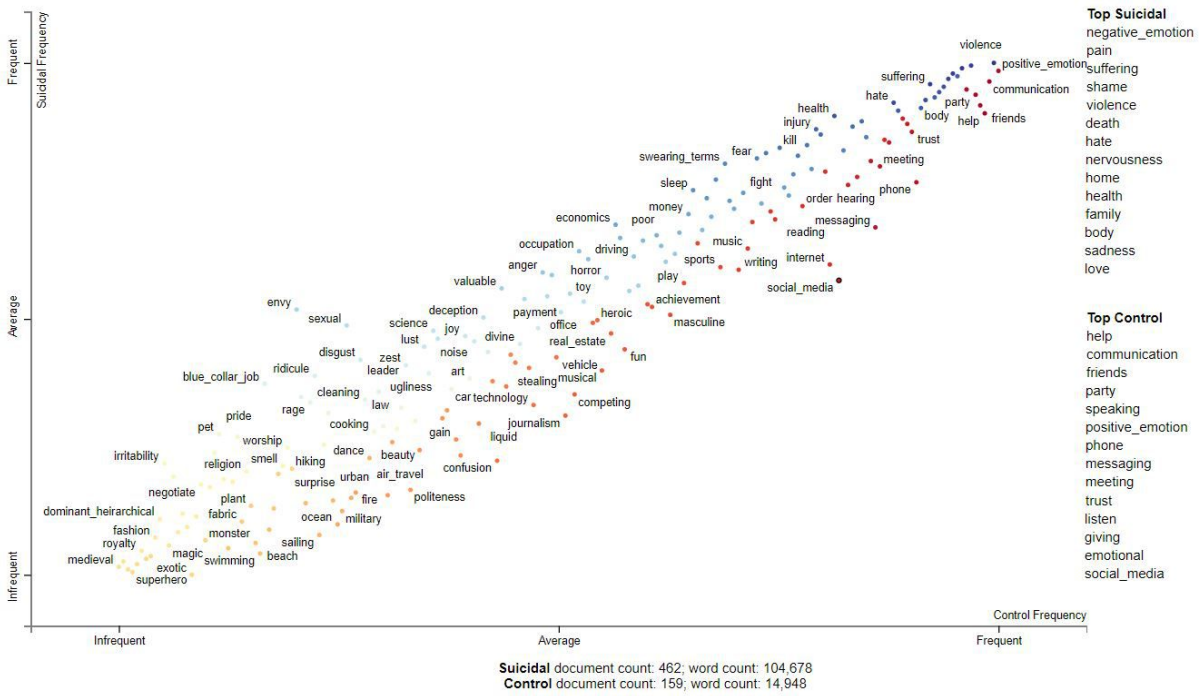

Figure 5.1. S1 - Topic modeling example

.

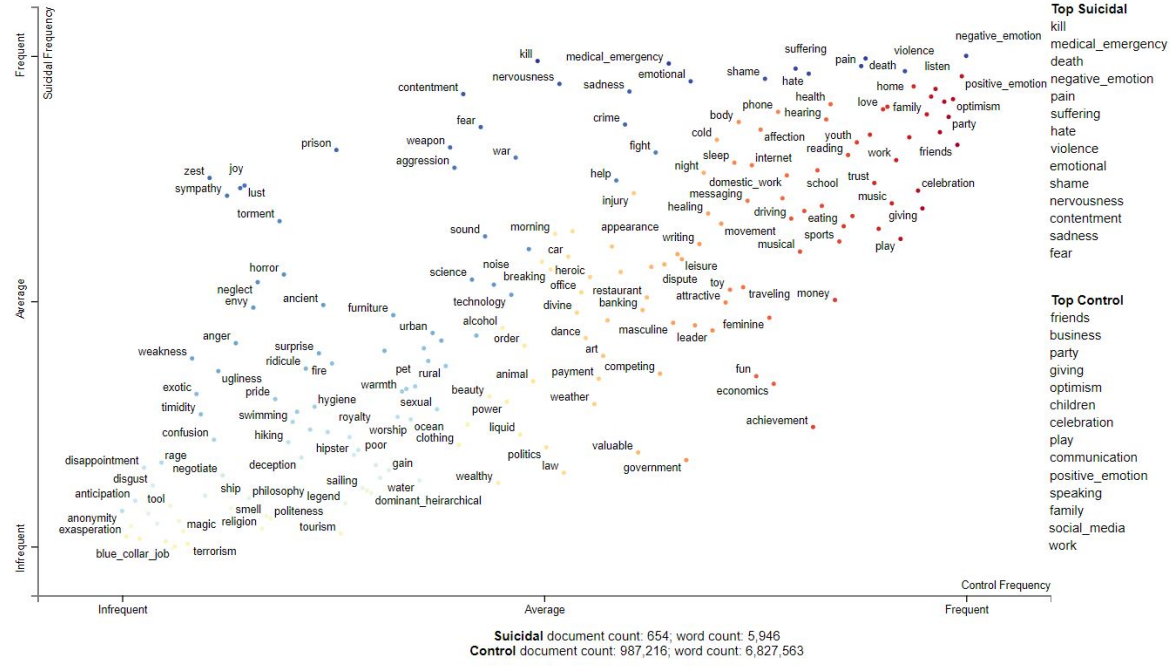

Figure 5.2. S2 - Topic modeling example

.

5.3 Results and Discussion

In this section, we evaluate four traditional and four deep learning algorithms for detecting users with suicide-ideation based on their posts on Reddit. We use 5-fold cross-validation with stratified sampling to evaluate the performance of each classifier in terms of accuracy, precision, recall and F1-score evaluation metrics.

5.3.1 The Results of the Traditional Classifiers

We used the following traditional classifiers: Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost).

The combination of each set of features was experimented and Table 5.2 shows the most important feature sets using different traditional machine learning algorithms for flagged users identification on the $\mathcal{S}1$ dataset. $\mathcal{S}1$ dataset contains the CLPSych 2019 training and test data combined together. We did this in order to be able to train a good classifier to run on the population-level data. And, as mentioned earlier, we report the results using 5-fold cross- validation.

The best features used were the linguistic (LIWC) and emotion features. Among these classifiers, XGBoost model achieved the best results with F1-score of 0.922 using LIWC and emotion features, while SVM achieved the best recall score of 0.971 using sentiment features.

5.3.2 Deep Learning Models

Deep learning models take the word-embedding matrix as the input of each model, the vocabulary contains 119,626 unique words, from which we selected the most frequent 50,000 words. The output is a single sigmoid activation function. The layers in between are sequential and fully connected; a dropout layer of probability 0.3 is used to avoid over-fitting. For the base model, we trained a 300-dimensional word embedding layer on the SuicideWatch subreddit posts then passed it to a bidirectional Gated Recurrent Unit layers (GRU). We use global max pooling with ReLu activation layer. And since we have a small dataset, we employed transfer learning model to enrich the training models as follows:

Table 5.2. Summary of the features used for predictive models for SI on the dataset $\mathcal{S}1$ with the best results achieved

Feature Set	**Classifier**	**Acc.%**	**Prec.**	**Recall**	**F1-score**
Stat.	SVM	80.8	0.844	0.911	0.875
	LR	80.6	0.844	0.907	0.874
	RF	77.6	0.793	0.946	0.862
	XGBoost	78.7	0.822	0.912	0.864
LIWC.	SVM	87.1	**0.907**	0.920	0.913
	LR	87.2	0.904	0.927	0.915
	RF	86.7	0.886	0.941	0.913
	XGBoost	86.1	0.894	0.924	0.908
Sent	SVM	77.1	0.777	**0.971**	0.862
	LR	76.9	0.782	0.956	0.859
	RF	77.1	0.800	0.921	0.855
	XGBoost	72.7	0.784	0.874	0.825
TopicModeling	SVM	80.5	0.837	0.917	0.874
	LR	80.6	0.839	0.917	0.875
	RF	77.6	0.790	0.951	0.862
	XGBoost	75.8	0.812	0.882	0.844
Doc2Vec	SVM	73.9	0.815	0.843	0.827
	LR	75.8	0.815	0.877	0.843
	RF	77.4	0.802	0.927	0.858
	XGBoost	77.7	0.823	0.892	0.855
tf-idf-Doc2Vec	SVM	72.4	0.805	0.832	0.817
	LR	74.8	0.812	0.862	0.835
	RF	77.4	0.804	0.922	0.858
	XGBoost	78.2	0.812	0.921	0.862
Sent+LIWC	SVM	84.5	0.896	0.896	0.895
	LR	86.6	0.900	0.922	0.911
	RF	87.0	0.886	0.948	0.915
	XGBoost	**88.1**	0.903	0.942	**0.922**

- BiLSTM_fastText: We used the pre-trained 300-dimentional fastText based on Common Crawl[3] as the word embedding layer, followed by a bidirectional Long Short Term Memory (LSTM). We used max pooling and average pooling to select the most representative features.

- CNN_fastText_LIWC: Inspired by the work of Zhou et al. (2016), we applied 2D convolutions and 2D pooling instead of 1D max pooling to sample more meaningful information from fastText word embedding layer. The convolution layer have 32 filters with the sizes of 2, 3 and 5. LIWC features were used in conjunction with the concatenated output of the max pooled word embedding convolution layers. Finally the decision is made by sigmoid function as shown in Figure 5.3.

- BertForSequenceClassification: Currently, BERT is the most popular NLP approach to transfer learning (Devlin et al., 2018). Using Huggingface[4] abstraction, we fine-turned the pre-trained BERT model using the Sentence Pair Classification model (BertForSequenceClassification) to obtain an F1-score 0.926 and an accuracy of 89.4%.

Model	**Acc.%**	**Prec.**	**Recall**	**F1-score**
Related Work (Mohammadi et al., 2019)	-	-	-	0.922
Bi-GRU (Baseline)	85.6	0.869	0.954	0.908
Bi-LSTM+fastText	85.8	0.877	0.943	0.907
CNN_fastText+LIWC	87.0	0.879	**0.959**	0.915
BertForSequenceClassification	**89.4**	**0.900**	0.955	**0.926**

Table 5.3. Deep learning models on $\mathcal{S}1$ dataset using 5-fold cross-validation

Table 5.3 shows different architectures for deep learning models that achieved comparable or slightly better F1-scores on the $\mathcal{S}1$ test dataset (0.926). The F1-scores that we obtained are comparable with the "(flagged) F1-score" of 0.922, the maximum F1-score achieved by the CLaC team in the CLPsych 2019 shared task (Zirikly et al., 2019).

[3]https://fasttext.cc

[4]https://huggingface.co/transformers/model_doc/bert.html

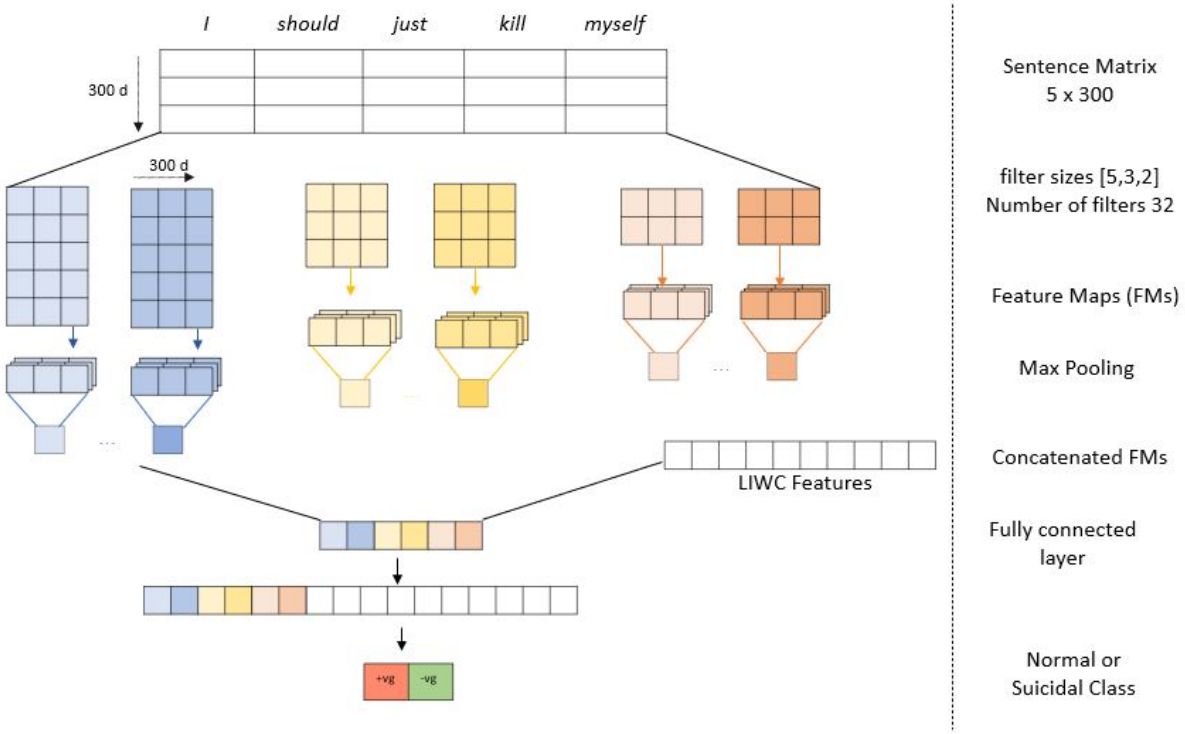

Figure 5.3. CNN_fastText_LIWC architecture inspired by (Kim, 2014)

5.4 Suicide Ideation Detection at Population Level

When restricted to limited training data and a small number of labeled users, using a good features set and a robust classifier may lead to competitive results. Nevertheless, deep learning architectures scored better accuracy, recall and F1-score than traditional classifiers. We tested the "best" traditional and deep learning models on the population level. CNN-fastText+LIWC has the best recall (0.959) and F1-score (0.915) on $\mathcal{S}1$. For traditional models we selected XGBoost fed with LIWC and sentiment features that scored a recall of 0.942 and F1-score of 0.922. We applied both systems on randomly selected users from $\mathcal{S}2$ with an equal number of suicidal and control users, and the results showed an accuracy of 83.5% and F1-score of 0.791 for XGBoost model and a recall of 0.881 and F1-score of 0.936 for CNN-fastText+LIWC.

Subsequently, we used the tuned models on the Canadian dataset $\mathcal{P}1$ for SI prediction. We utilized the data from the 2015-2016 Canadian Community Health Survey–Annual Component on the adult population of Canada's provinces and territories *(n = 109,659)*. The survey data is based on variable SUI_010 asking respondents regarding considering suicide during the last 12 months: *"Has this happened in the past 12 months?"*. The question was not asked in proxy interviews, and 21.4% answered *yes*. In addition, based on Statistics Canada[5] an estimate of the total number of people who reported serious suicidal thoughts is 3,396,700 persons based on data from 2015, i.e., 9.514% of the total population. CNN-fastText+LIWC model - for the same year - predicted 3,938 from social media users with suicidal thoughts, i.e., 10.6% of the population sample $\mathcal{P}1$ dataset, while XGBoost model predicted 2,583 suicidal users (7.3%).

In the following sections, we draw the differences between the traditional and deep learning predictions on the population-level:

5.4.1 Traditional: XGBoost Classifier

A Pearson correlation analysis showed 0.61 correlation between the predicted number of users based on XGBoost and the positive respondents to the CCHS suicide attempt survey.

Table 5.4 and Figure 5.4 show the XGBoost model's predictions and the actual values.

[5]Statistics Canada. Table 13-10-0098-01 Mental health characteristics and suicidal thoughts

Table 5.4. XGBoost model estimation number and percentage versus reported suicidal thoughts percentage (by age group) based on CCHS (2015-2016) per province

Provinces	**<25**	**25-34**	**35-44**	**45-54**	**55-64**	**≥ 65**	**Predicted%**	**Actual**
AB	98	41	61	44	6	14	13.46	11.62
BC	73	81	107	194	115	47	12.27	10.08
MN	2	3	6	4	2	1	9.42	8.85
NB	39	9	205	2	13	12	13.49	11.18
NL	25	45	55	106	104	6	1.99	7.43
NS	5	18	12	19	22	2	9.69	9.51
ON	88	62	103	253	175	77	10.19	8.28
PE	8	3	4	1	6	4	5.73	7.07
SK	71	9	9	21	75	16	29.13	10.22

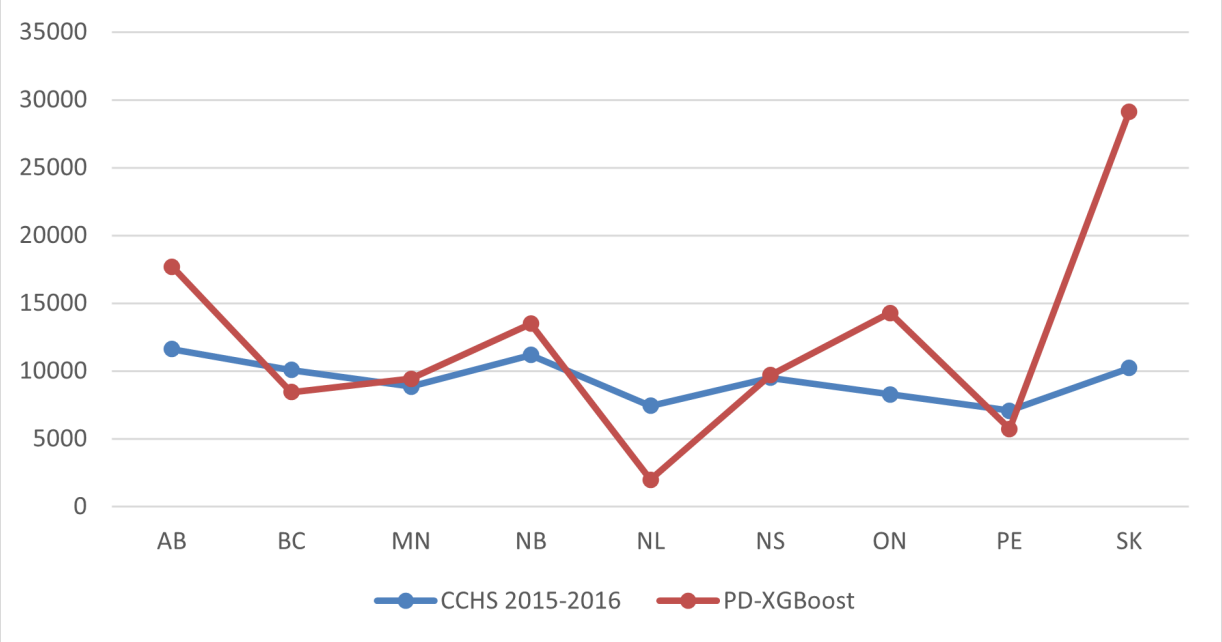

Figure 5.4. Predicted SI using XGBoost model versus actual statistics

5.4.2 Deep Learning: fastText+LIWC Classifier

Table 5.5 shows the Pearson linear correlation (ρ) of the demographic characteristic between CCHS 2015-2016 and the estimations using the "CNN-fastText+LIWC" on $\mathcal{P}1$ dataset. Figure 5.5 shows the comparison between the estimated and the actual suicide ideation percentage during 2015 for the nine provinces and two territories, according to CCHS 2015-2016.

Table 5.5. Pearson linear correlation between CCHS 2015-2016 and the prediction on $\mathcal{P}1$ dataset using CNN-fastText+LIWC model

	Pearson Correlation
Sex	
Female	0.99
Male	0.95
Age	
<25	0.99
25-34	0.87
35-44	0.94
45-54	0.97
55-64	0.97
≥65	0.99
Province	
Alberta	0.84
British Columbia	0.92
Manitoba	0.98
New Brunswick	0.83
Newfoundland and Labrador	0.82
Northwest Territories	0.63
Nova Scotia	0.65
Ontario	0.90
Prince Edward Island	0.75
Saskatchewan	0.62
Yukon	0.69

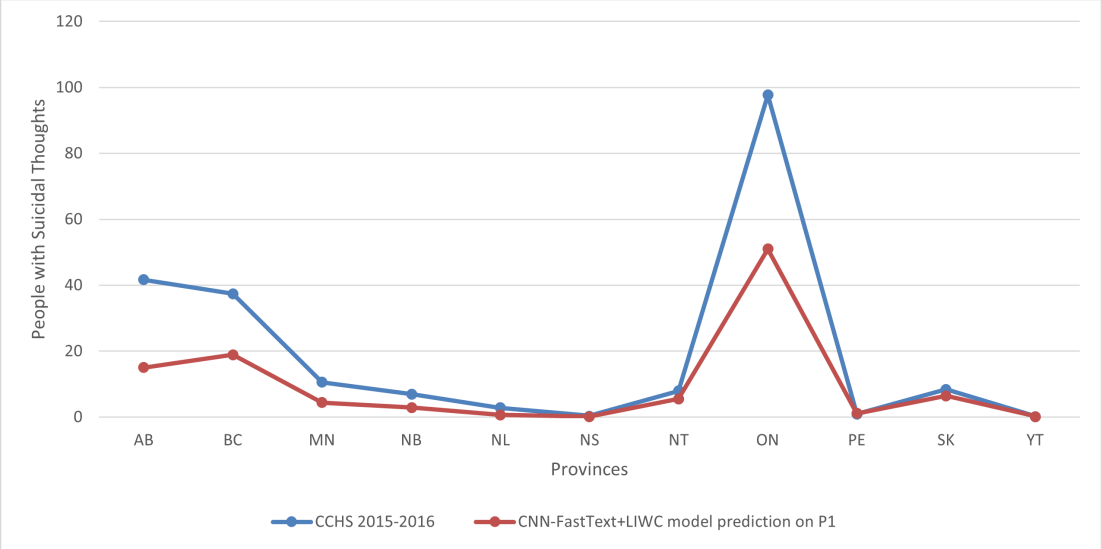

Figure 5.5. Predicted SI CNN-fastText+LIWC versus actual statistics

Table 5.6 and Figure 5.8 shows the distribution of age demographics within 9 of the Canadian provinces and two territories and the estimated suicide ideation on P1 dataset. Similarly, Table 5.7 and Figure 5.6 shows the distribution of males and females with suicidal thoughts based on CCHS and the predicted suicide ideation on P1 dataset. This adheres to the fact that females account for more suicidal thoughts among adults, and males account for more suicide execution[6].

5.5 Summary

In this chapter, we applied traditional machine learning algorithms with feature engineering to predict suicide ideation at the user level. We compared the results with four deep learning algorithms, then we applied the best performing models from each (XGBoost with LIWC and sentiment features, and CNN-fastText with LIWC features) on the population-representative dataset. We compared the two models and CNN showed stronger correlation with the CCHS statistics data. The results showed a strong positive correlation between the predicted suicide ideation and the available statistics for 2015 at province level (0.62 and 0.98).

[6]based on Canada's Public Health Agency data

Table 5.6. CNN-fastText+LIWC model estimation versus reported suicidal thoughts percentage (by age group) based on CCHS (2015-2016) per province

Pr.	**Predictions on P1 Dataset %**						**Census 2015-2016 Survey%**					
	<25	25-34	35-44	45-54	55-64	≥ 65	<25	25-34	35-44	45-54	55-64	≥65
AB	5.54	2.41	1.78	1.63	1.45	1.30	5.64	4.46	3.55	2.47	2.35	0.90
BC	7.16	3.73	1.85	1.57	2.01	1.50	5.73	3.17	2.68	2.59	2.49	0.76
MN	1.35	1.07	0.48	0.53	0.38	0.33	1.93	1.17	0.42	0.63	0.47	0.30
NB	0.99	0.71	0.18	0.33	0.25	0.28	1.10	0.61	0.51	0.56	0.36	0.10
NL	0.30	0.08	0.05	0.05	0.03	0.15	0.61	0.26	0.13	0.14	0.11	0.07
NT	0.13	0.03	0.03	0.03	0.03	0.00	0.05	0.06	0.01	0.02	0.03	0.00
NS	2.16	1.07	0.48	0.38	0.46	0.61	0.98	0.93	0.27	0.79	0.47	0.25
ON	21.15	3.33	2.16	11.58	5.05	4.57	16.37	7.39	5.80	8.70	4.45	2.72
PE	0.43	0.15	0.30	0.05	0.03	0.08	0.12	0.04	0.12	0.05	0.08	0.01
SK	1.90	1.04	0.86	1.42	0.46	0.33	1.16	0.93	1.06	0.34	0.31	0.10
YT	0.03	0.03	0.10	0.05	0.03	0.03	0.01	0.04	0.05	0.03	0.01	0.00

Table 5.7. CNN-fastText+LIWC model estimation versus actual males and females with suicidal thoughts based on CCHS (2015-2016) per province

Province	**Model prediction**				**Census 2015-2016 Survey**			
	Females	F%	Males	M%	Females	F%	Males	M%
AB	381	0.5	174	1.5	61,159	2.5	41,983	1.7
BC	443	0.6	259	2.2	47,121	1.9	45,559	1.8
MN	92	0.1	71	0.3	16,306	0.7	9,878	0.4
NB	64	0.0	44	0.2	8,905	0.4	8,384	0.3
NL	17	0.0	9	0.1	3,774	0.2	3,265	0.1
NS	7	0.0	2	0.0	496	0.5	470	0.3
NT	142	0.2	61	0.5	11,192	0	8,431	0
ON	1,075	1.1	809	3.9	150,987	6.1	90,865	3.7
PE	23	0.0	18	0.1	1,113	0	1,129	0
SK	82	0.1	155	0.4	11,791	0.5	8,991	0.4
YT	3	0.0	7	0.0	165	0	528	0

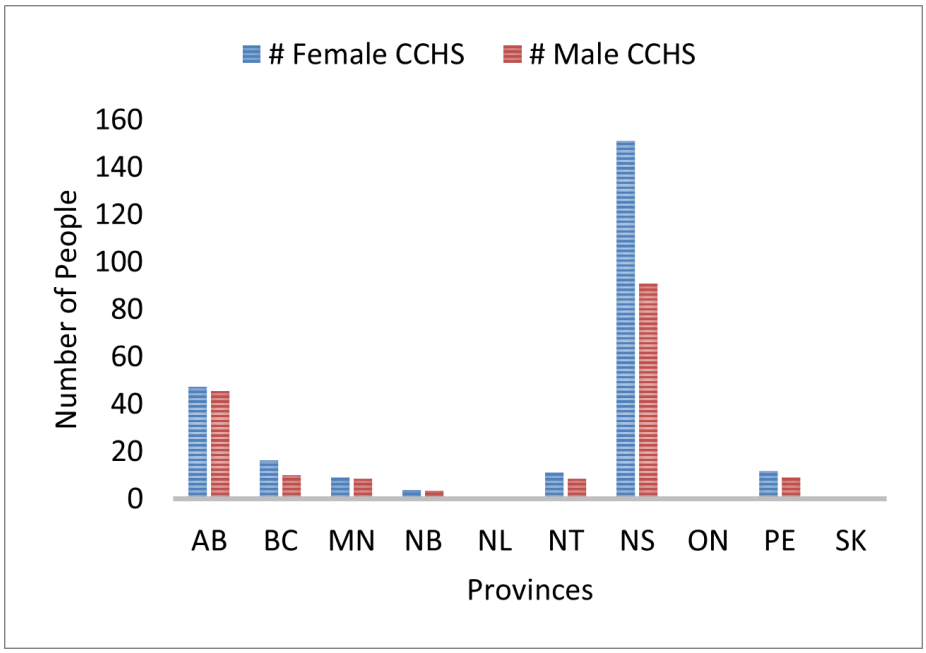

Figure 5.6. SI among males and females in Canada's provinces based on CCHS 2015-2016 statistics (per 1,000 population)

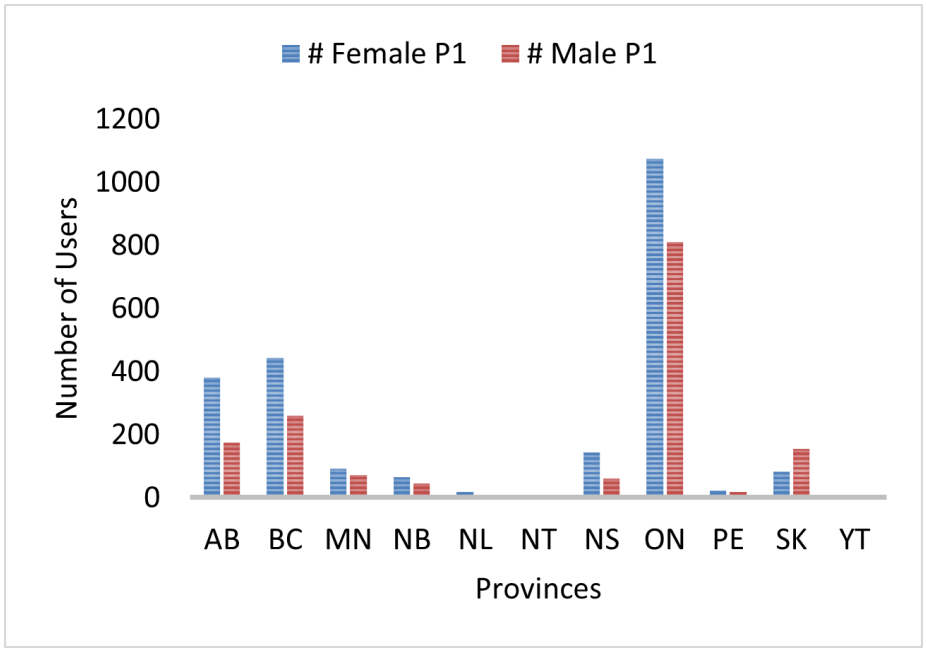

Figure 5.7. Estimated SI among females and males on P1 dataset based on CNN-fastText+LIWC model

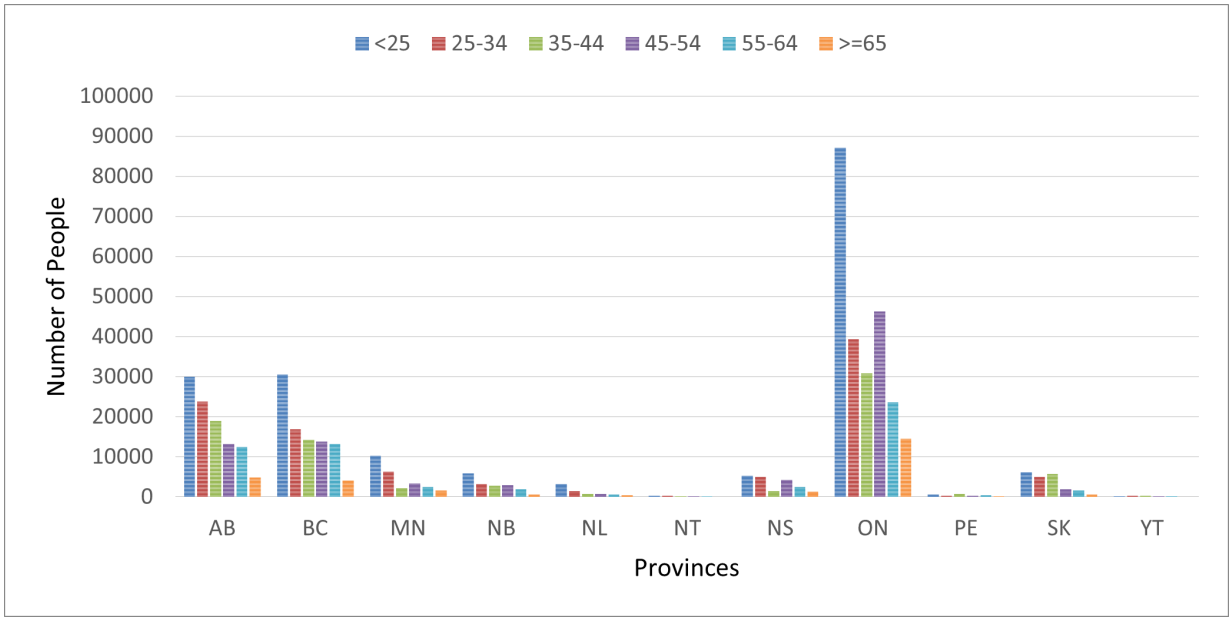

Figure 5.8. SI among different age groups based on CCHS 2015-2016 statistics

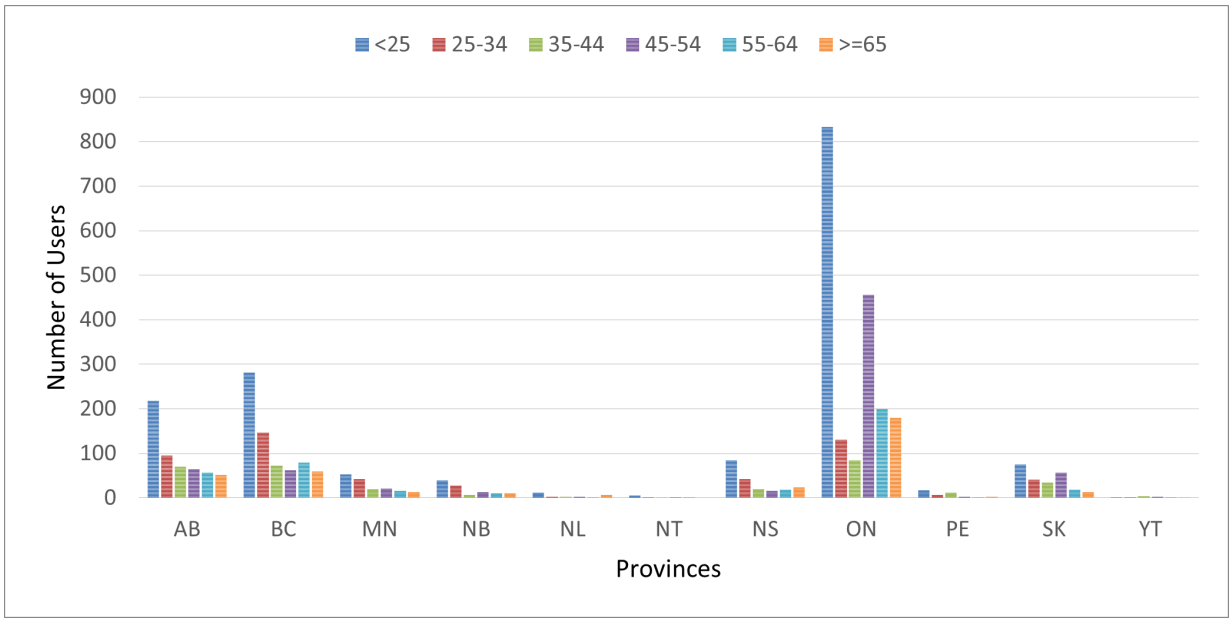

Figure 5.9. Estimated SI among different age groups on P1 dataset

Chapter 6

Conclusion and FutureWork

6.1 Conclusion

Social media analysis can be used to uncover depression trends globally and extract information for diagnosis and prevention. This thesis addressed depression and suicide ideation classification from social media texts for public health surveillance.

We used a well-labeled dataset based on Twitter for depression classification and another labeled dataset based on Reddit for suicide ideation classification. Both datasets are parts of the shared tasks organized by CLPSych. Then we tested different models on Twitter-based datasets that resemble the population dataset under study and is labeled using self-disclosure. The test datasets contrast the training datasets in many aspects, including volume, labeling techniques, and anonymization. The models were analyzed using multiple features and evaluated based on different metrics, mainly F1-score and recall, and the best model was chosen for depression and suicide ideation population estimation. We established the standards and recommendations when collecting a population-level representative dataset from social media data. The correlation between the prediction and actual statistics is $> +0.5$, and therefore we can say that we proved our hypothesis from Section 1.4. We can conclude that recent ML and NLP developments for mental health studies have showed promising results and can remarkably increase mental illness identification at the post level, at user level, and even at population level.

The results of this analysis might be helpful for population health planning purposes. To our knowledge, this is the first study to utilize social media data for depression and suicide ideation to provide an estimate of these two illnesses within the Canadian population and relate it to the CCHS.

6.2 Limitations and Challenges

There are still many challenges and questions that need to be addressed to exploit the opportunities for using social media data to predict mental health issues within the population. This section briefly discusses some of the limitations and challenges that arise in this endeavor, offering some recommendations to overcome such barriers.

6.2.1 Availability and Correctness of Social Media Data

One of the major issues in using NLP and ML tools for public mental health prediction is the availability of correctly-labeled data. Social media provides more than ten times the volume of data collected by country-wide surveys; however collecting relevant posting is challenging because of semantic heterogeneity and diverse writing styles. Moreover, labeling a good sample is time-consuming and requires professional resources and a degree of consensus. Using self-reports as a way to gather signs of mental disorders within posts is the easiest, but it may not be considered as clinical ground truth since there is no way to verify an actual diagnosis, nor any way to determine that a control instance is not positive for the conditions. Semi-automatic labeling techniques could be adopted to label the data, as previously stated, but more accurate labeling techniques need to be researched to minimize human interference. Another approach is to dig into the users' posts looking for symptoms and finding psychiatric questionnaires answers to label the data automatically.

6.2.2 Is Social Media Representative of the General Population?

A fundamental limitation in this field is that social media users are not representative of the general population (Mellon and Prosser, 2017). Social media is biased in different ways. For example, most Twitter users are of an age between 18 and 34 years old, and

most Facebook users are female[1]. Besides, there is a variance in posting using users' geo-location based on age, gender, ethnic, and socio-demographic groups (Rzeszewski and Beluch, 2017). Conventional methods suffer from excluding significant clusters within the population that could be more prone to mental illness such as Aboriginal people and members of the Canadian Forces. Other attributes can negatively affect the presence of under-served populations and minority groups in social media, such as lack of education or Internet access (Denecke et al., 2015).

These limitations can be resolved using different techniques such as user filtering (Filho et al., 2015; Yang et al., 2015), characterizing demographic attributes (Chen et al., 2015) and user sampling (White et al., 2012; Aghababaei and Makrehchi, 2017). User attributes, such as gender, age, ethnicity, and location can be predicted from the messages posted by the user on social media (Farzindar and Inkpen, 2018). Similar studies can be conducted on children and teens related data taking under consideration the Children's Online Privacy and Protection Act (COPPA). Although using social media contents shows promise in mental health detection within the population, significant research on representative samples is necessary to validate the results and strengthen the models.

6.2.3 Importance of Identifying Risk Factors

Specific risk factors include age, sex, substance abuse, certain medications, chronic diseases, family, workplace, genetics, illness, and major life event like marriage, divorce, death, or abuse. More research is needed to better understand the influence of features for identifying such risk factors and consequently, efficiently classifying social media users with signs of mental illness to support mental health monitoring at the population-level. ML offers an opportunity to delve into complex interactions between risk factors.

6.2.4 Ethics

There are ethical challenges in using social media as a source for NLP and ML. Conway (2014) provided a taxonomy of ethical principles on the use of Twitter in public health research based on a review of the literature. These principles can be implemented among

[1]https://insightswest.com/wp-content/uploads/2017/10/Rep_IW_CDNSocialMediaMonitor_Oct2017.pdf

all social media platforms. Researchers face important challenges to ensure the privacy of social media data (Conway et al., 2016; Gruebner et al., 2017b). Although some of the data is publicly available, the problem becomes more complicated when personal attributes can be predicted, and the identity of particular users can be revealed. There is an elaborated discussion of ethics, particularly in public health research such as (Mckee, 2013; Mikal et al., 2016; Denecke et al., 2015; Valdez and Keim-Malpass, 2019). Based on their experience in the domain, Benton et al. (2017a) developed practical protocols to guide NLP research using social media data from a healthcare perspective. Their guidelines recommend that researchers need to acquire an ethical approval or exemption from their Institutional Review Board (IRB). Researchers also need to obtain informed consent when possible and protect and anonymize sensitive data when used in presentations or analysis. Besides, they need to be vigilant when linking data across sites is necessary. Finally, when sharing their data, they need to ensure that other researchers respect ethical and privacy concerns.

In general, there is an agreement that researchers can use publicly available data for health monitoring, but preserving the confidentiality of social media users is a must. Predicting the clusters vulnerable to mental disorders is one of the steps in protective medicine. After identifying such groups, there are practical steps that need to be considered from responsible parties to collaborate for disease control, treatment, and prevention. These steps may require informed consent and acceptance of such interventions in the target population or required government-related programs to address specific clusters and provide the appropriate help.

6.3 Future Work

6.3.1 Fine-grained Categories:

In order to better serve communities, the models could be trained on more fine-grained categories to classify people at high, medium, low risk, or no risk. The CLPsych2019 data contains the four labels. Thus, we plan to use transfer learning, by utilizing the word embedding trained on mental health corpus such as Depression Specific Word Embedding (DSE) and Adjusted Twitter Word Embedding (ATE) (Farruque et al., 2020) and compare them with the fine-tuned BERT model or GPT-3 model with few-shot learning.

6.3.2 Ensemble ML Algorithm for Population Prediction:

Different decision fusion methods can be tested to integrates the conclusions of several classifiers into a single conclusion. For example, the current methodology can be extended by using multiple ML algorithms to predict users as risk. Each model's output can be a set of users with a weighted risk vector. On top of them, an ensemble model can cluster the users into different levels of risk based on a voting rule as illustrated in Figure 6.1.

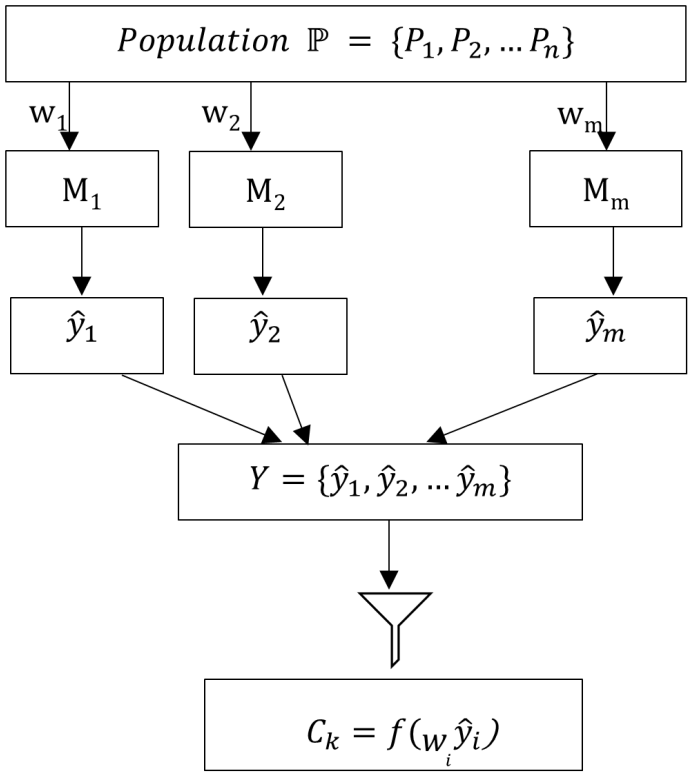

Figure 6.1. Ensemble users-at-risk prediction based on m Models(M) for population P, where $\hat{y}_i$ is the resultant prediction on M_i and w_i is the weight for each model based on its performance on the test dataset

6.3.3 Multi-language Social Media Analysis

CCSH (2015-2016) shows that 98.2% of Canadians speak either French or English or both, but 19.5% speak only French. Besides, 22.6% have mother tongues other than English or French. The familiarity of the language affects its use on social media. To extend this work, we plan to include French language in our predictive model and use multilingual models such as mBERT or monolingual models such as CamemBERT[2] - a state-of-the-art French language model built on the RoBERTa architecture - and FlauBERT (Le et al., 2020).

6.3.4 Depression Symptoms Detection

Utilizing the dictionary and lexicon of the nine major signs of depression and using few-shots learning on the dataset provided by the eRisk 2020[3] shared task 2 for measuring the severity of the signs of depression. We plan to answer the PHQ-9 questionnaire based on the users' posts[4].

The posts of 20 users along with the answered questionnaire are provided to the shared task participants. The users' posts can be analyzed to answer the PHQ-9 questionnaire that demonstrated acceptable reliability for evaluating the severity of the depressive symptoms (Maroufizadeh et al., 2019).

The proposed strategy is illustrated in Figure 6.2 and described as follows:

- Each user is represented by a series of posts $T_1, T_2, ...T_n$ that are aggregated for two weeks period each.
- Each tweet T_i is represented as a word embedding vector and set of tweets is forming a document vector D_i representing the tweets of $period_i$.
- The semantic similarity between D_i and the symptoms lexicons and the DBI questions and answers text is calculated for each symptom.

[2]https://camembert-model.fr/

[3]https://early.irlab.org/

[4]We have conducted many experiments to filter out tweets based on this method, we were able to match the social media tweets with PHQ-9 but more experiments in this area need to be done to be able to select the "best" periods of tweets that represent the social media user.

- The user will be represented by the average of the maximum 3 consecutive periods as follows: $\max(x_{k-1}, x_k, x_{k+1})$ where $x_k = avg(\forall_j T_i(S_j))$
- Based on the most representative tweets of the user, the symptoms array $S(U_i)$ is formed as an average of each symptom for the three weeks period and concatenated with the word embedding of the tweets for that period to be the input layer of an LSTM network.

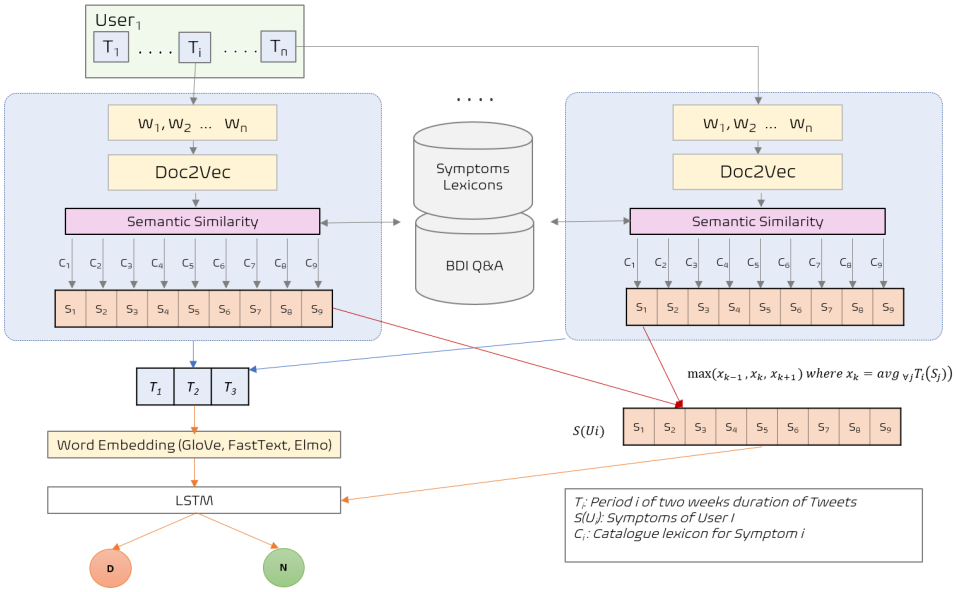

Figure 6.2. Depression symptoms estimator

6.3.5 Suicide Ideation Enhancment

Enhance the performance of the suicide ideation model using different methods with a main focus on COVID-19 implications. Knowing that suicide is 25 times more likely in those suffering from depression[5], one method to enhance suicide ideation prediction is by adding a depression indicator - estimated from depression models. Another method could use letter bi-grams so that every text will be represented as a 28x28 matrix, every entry x_i reflects the weight of a particular letter combination in the normalized text as an input to a deep neural network such as CNN.

[5]https://www.suicideinfo.ca/resource/depression-suicide-prevention/

APPENDICES

Appendix A

Experimental Details

A.1 LIWC Categories

The Table A.1 shows the 93 features produced by the LIWC tool for two examples based on $\mathcal{D}1$ dataset. The first example is the sentence: "I got my iPhone back yay !!!x", and the second example is the sentence "Hurt people hurt people....that is all"[1].

Table A.1. The properties of LIWC-2015 (with the output of the two examples)

LIWC category	**LIWC label**	**Example 1**	**Example 2**
Word Count	WC	7	7
Summary Variable			
Analytical Thinking	Analytic	5.89	29.3
Clout	Clout	1	99
Authentic	Authentic	99	1
Emotional Tone	Tone	99	1
Language Metrics			
Words per sentence1	WPS	7	7
Words>6 letters	Sixltr	0	0
Dictionary words	Dic	71.43	100
Function Words	function	42.86	28.57
Total pronouns	pronoun	28.57	14.29
Personal pronouns	ppron	28.57	0

[1]Some words were deleted to maintain users' privacy.

LIWC-2015 properties, labels and examples (Cont.)

LIWC property	LIWC label	Example 1	Example 2
Language Metrics (Cont.)			
1st pers singular	i	28.57	0
1st pers plural	we	0	0
2nd person	you	0	0
3rd pers singular	shehe	0	0
3rd pers plural	they	0	0
Impersonal pronouns	ipron	0	14.29
Articles	article	0	0
Prepositions	prep	0	0
Auxiliary verbs	auxverb	0	14.29
Common adverbs	adverb	14.29	0
Conjunctions	conj	0	0
Negations	negate	0	0
Grammar Other			
Regular verbs	verb	14.29	14.29
Adjectives	adj	0	0
Comparatives	compare	0	0
Interrogatives	interrog	0	0
Numbers	number	0	0
Quantifiers	quant	0	14.29
Affect Words	affect	14.29	28.57
Positive emotion	posemo	14.29	0
Negative emotion	negemo	0	28.57
Anxiety	anx	0	0
Anger	anger	0	0
Sadness	sad	0	28.57
Social Words	social	0	28.57
Family	family	0	0
Friends	friend	0	0
Female referents	female	0	0
Male referents	male	0	0
Cognitive Processes	cogproc	0	14.29
Insight	insight	0	0
Cause	cause	0	0
Discrepancies	discrep	0	0
Tentativeness	tentat	0	0
Certainty	certain	0	14.29
Differentiation	differ	0	0
Perpetual Processes	percept	0	28.57
Seeing	see	0	0
Hearing	hear	0	0

LIWC-2015 properties, labels and examples (Cont.)

LIWC category	LIWC label	Example 1	Example 2
Grammar Other (Cont.)			
Feeling	feel	0	28.57
Biological Processes	bio	0	0
Body	body	0	0
Health/illness	health	0	0
Sexuality	sexual	0	0
Ingesting	ingest	0	0
Core Drives and Needs	drives	14.29	0
Affiliation	affiliation	0	0
Achievement	achieve	0	0
Power	power	0	0
Reward focus	reward	14.29	0
Risk/prevention focus	risk	0	0
Time Orientation			
Past focus	focuspast	14.29	0
Present focus	focuspresent	0	14.29
Future focus	focusfuture	0	0
Relativity	relativ	14.29	0
Motion	motion	0	0
Space	space	14.29	0
Time	time	14.29	0
Personal Concerns			
Work	work	0	0
Leisure	leisure	0	0
Home	home	0	0
Money	money	0	0
Religion	relig	0	0
Death	death	0	0
Informal Speech	informal	14.29	0
Swear words	swear	0	0
Netspeak	netspeak	0	0
Assent	assent	14.29	0
Nonfluencies	nonflu	0	0
Fillers	filler	0	0
All Punctuation	AllPunc	42.86	57.14
Periods	Period	0	57.14
Commas	Comma	0	0
Colons	Colon	0	0
Semicolons	SemiC	0	0

LIWC-2015 properties, labels and examples (Cont.)

LIWC category	LIWC label	Example 1	Example 2
Personal Concerns (Cont.)			
Question marks	QMark	0	0
Exclamation marks	Exclam	42.86	0
Dashes	Dash	0	0
Quotation marks	Quote	0	0
Apostrophes	Apostro	0	0
Parentheses (pairs)	Parenth	0	0
Other punctuation	OtherP	0	0

A.2 Depression Statistical Significance Test

Statistical significance test was conducted on D1 test using the 5x2 cross-validation paired t-test from MLxtend[2]. The p-value was calculated to validate that there is a statistical significance difference between the different machine learning algorithms as shown in Table A.2 where (*) indicates that models with $p\text{-}value < 0.05$. Figure A.1 shows the performance of SVM, LR, RF, GBDT, and XGBoost algorithms using Datapane python library[3].

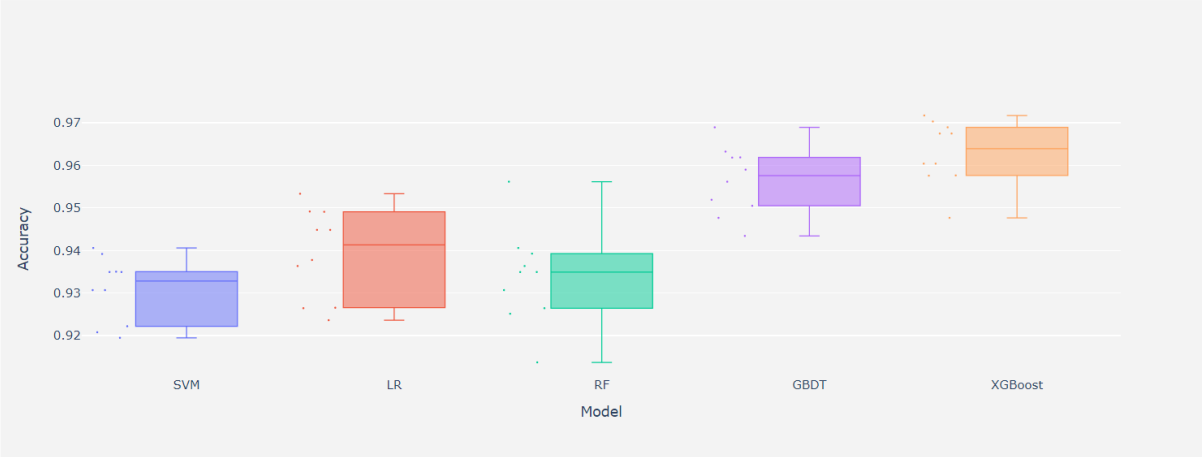

Figure A.1. Performance of different models using 5x2-fold cross-validation

[2]http://rasbt.github.io/mlxtend/user_guide/evaluate/paired_ttest_5x2cv/
[3]https://github.com/datapane

Table A.2. Comparison of various ML algorithms using different features on $\mathcal{D}1$ dataset (5x2 cross-validation, (*) indicates statistical significance with the other models ($p < 0.05$)

Features	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Basic Statistics	SVM	0.708	0.666	0.654	0.659
	LR	0.709	0.664	0.661	0.662
	RF	*0.727	0.691	0.663	**0.676**
	GBDT	*0.725	0.686	0.667	**0.676**
	XGBoost	*0.715	0.675	0.656	0.665
Tf-idf	SVM	0.894	0.883	0.869	0.876
	LR	0.894	0.883	0.869	0.875
	RF	0.902	0.885	0.890	0.887
	GBDT	*0.903	0.892	0.882	0.887
	XGBoost	*0.912	0.897	0.899	**0.898**
POS	SVM	0.686	0.608	0.767	0.678
	LR	*0.685	0.609	0.755	0.674
	RF	*0.741	0.687	0.735	0.710
	GBDT	*0.762	0.731	0.710	**0.720**
	XGBoost	*0.752	0.713	0.713	0.713
Topics	SVM	0.749	0.697	0.736	0.716
	LR	*0.749	0.703	0.724	0.713
	RF	*0.764	0.707	0.775	**0.739**
	GBDT	*0.743	0.699	0.710	0.704
	XGBoost	*0.754	0.711	0.726	0.718
LIWC	SVM	0.815	0.813	0.742	0.776
	LR	0.819	0.814	0.752	0.782
	RF	0.814	0.803	0.753	0.777
	GBDT	*0.843	0.838	0.789	**0.812**
	XGBoost	*0.833	0.816	0.791	0.803
Sentiment	SVM	0.606	0.585	0.295	0.391
	LR	0.602	0.570	0.307	0.399
	RF	*0.652	0.596	0.611	0.602
	GBDT	*0.647	0.594	0.574	0.584
	XGBoost	0.656	0.599	0.614	**0.605**
All Features	SVM	91.2	89.1	0.908	0.899
	LR	92.7	0.911	0.921	0.916
	RF	94.4	0.938	0.932	0.935
	GBDT	***96.0**	**0.964**	0.958	**0.961**
	XGBoost	*96.4	0.956	**0.960**	0.958

A.3 Full BERT Model

```
BertForSequenceClassification(
  (bert): BertModel(
    (embeddings): BertEmbeddings(
      (word_embeddings): Embedding(30522, 768, padding_idx=0)
      (position_embeddings): Embedding(512, 768)
      (token_type_embeddings): Embedding(2, 768)
      (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      (dropout): Dropout(p=0.1, inplace=False))
    (encoder): BertEncoder(
      (layer): ModuleList(
        (0): BertLayer(
          (attention): BertAttention(
            (self): BertSelfAttention(
              (query): Linear(in_features=768, out_features=768, bias=True)
              (key): Linear(in_features=768, out_features=768, bias=True)
              (value): Linear(in_features=768, out_features=768, bias=True)
              (dropout): Dropout(p=0.1, inplace=False))
            (output): BertSelfOutput(
              (dense): Linear(in_features=768, out_features=768, bias=True)
              (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
              (dropout): Dropout(p=0.1, inplace=False))
          )
          (intermediate): BertIntermediate(
            (dense): Linear(in_features=768, out_features=3072, bias=True))
          (output): BertOutput(
            (dense): Linear(in_features=3072, out_features=768, bias=True)
            (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
            (dropout): Dropout(p=0.1, inplace=False))
        )
        (1): BertLayer(
          (attention): BertAttention(
            (self): BertSelfAttention(
              (query): Linear(in_features=768, out_features=768, bias=True)
              (key): Linear(in_features=768, out_features=768, bias=True)
              (value): Linear(in_features=768, out_features=768, bias=True)
              (dropout): Dropout(p=0.1, inplace=False))
            (output): BertSelfOutput(
              (dense): Linear(in_features=768, out_features=768, bias=True)
              (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
              (dropout): Dropout(p=0.1, inplace=False))
          )
          (intermediate): BertIntermediate(
            (dense): Linear(in_features=768, out_features=3072, bias=True))
          (output): BertOutput(
            (dense): Linear(in_features=3072, out_features=768, bias=True)
            (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
            (dropout): Dropout(p=0.1, inplace=False))
```

```
)
(2): BertLayer(
  (attention): BertAttention(
    (self): BertSelfAttention(
      (query): Linear(in_features=768, out_features=768, bias=True)
      (key): Linear(in_features=768, out_features=768, bias=True)
      (value): Linear(in_features=768, out_features=768, bias=True)
      (dropout): Dropout(p=0.1, inplace=False))
    (output): BertSelfOutput(
      (dense): Linear(in_features=768, out_features=768, bias=True)
      (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      (dropout): Dropout(p=0.1, inplace=False))
  )
  (intermediate): BertIntermediate(
    (dense): Linear(in_features=768, out_features=3072, bias=True))
  (output): BertOutput(
    (dense): Linear(in_features=3072, out_features=768, bias=True)
    (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
    (dropout): Dropout(p=0.1, inplace=False))
)
(3): BertLayer(
  (attention): BertAttention(
    (self): BertSelfAttention(
      (query): Linear(in_features=768, out_features=768, bias=True)
      (key): Linear(in_features=768, out_features=768, bias=True)
      (value): Linear(in_features=768, out_features=768, bias=True)
      (dropout): Dropout(p=0.1, inplace=False))
    (output): BertSelfOutput(
      (dense): Linear(in_features=768, out_features=768, bias=True)
      (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      (dropout): Dropout(p=0.1, inplace=False))
  )
  (intermediate): BertIntermediate(
    (dense): Linear(in_features=768, out_features=3072, bias=True))
  (output): BertOutput(
    (dense): Linear(in_features=3072, out_features=768, bias=True)
    (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
    (dropout): Dropout(p=0.1, inplace=False))
)
(4): BertLayer(
  (attention): BertAttention(
    (self): BertSelfAttention(
      (query): Linear(in_features=768, out_features=768, bias=True)
      (key): Linear(in_features=768, out_features=768, bias=True)
      (value): Linear(in_features=768, out_features=768, bias=True)
      (dropout): Dropout(p=0.1, inplace=False))
    (output): BertSelfOutput(
      (dense): Linear(in_features=768, out_features=768, bias=True)
      (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
```

```
      (dropout): Dropout(p=0.1, inplace=False))
  )
  (intermediate): BertIntermediate(
    (dense): Linear(in_features=768, out_features=3072, bias=True))
  (output): BertOutput(
    (dense): Linear(in_features=3072, out_features=768, bias=True)
    (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
    (dropout): Dropout(p=0.1, inplace=False))
)
(5): BertLayer(
  (attention): BertAttention(
    (self): BertSelfAttention(
      (query): Linear(in_features=768, out_features=768, bias=True)
      (key): Linear(in_features=768, out_features=768, bias=True)
      (value): Linear(in_features=768, out_features=768, bias=True)
      (dropout): Dropout(p=0.1, inplace=False))
    (output): BertSelfOutput(
      (dense): Linear(in_features=768, out_features=768, bias=True)
      (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      (dropout): Dropout(p=0.1, inplace=False))
  )
  (intermediate): BertIntermediate(
    (dense): Linear(in_features=768, out_features=3072, bias=True))
  (output): BertOutput(
    (dense): Linear(in_features=3072, out_features=768, bias=True)
    (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
    (dropout): Dropout(p=0.1, inplace=False))
)
(6): BertLayer(
  (attention): BertAttention(
    (self): BertSelfAttention(
      (query): Linear(in_features=768, out_features=768, bias=True)
      (key): Linear(in_features=768, out_features=768, bias=True)
      (value): Linear(in_features=768, out_features=768, bias=True)
      (dropout): Dropout(p=0.1, inplace=False))
    (output): BertSelfOutput(
      (dense): Linear(in_features=768, out_features=768, bias=True)
      (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      (dropout): Dropout(p=0.1, inplace=False))
  )
  (intermediate): BertIntermediate(
    (dense): Linear(in_features=768, out_features=3072, bias=True))
  (output): BertOutput(
    (dense): Linear(in_features=3072, out_features=768, bias=True)
    (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
    (dropout): Dropout(p=0.1, inplace=False))
)
(7): BertLayer(
  (attention): BertAttention(
```

```
    (self): BertSelfAttention(
      (query): Linear(in_features=768, out_features=768, bias=True)
      (key): Linear(in_features=768, out_features=768, bias=True)
      (value): Linear(in_features=768, out_features=768, bias=True)
      (dropout): Dropout(p=0.1, inplace=False))
    (output): BertSelfOutput(
      (dense): Linear(in_features=768, out_features=768, bias=True)
      (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      (dropout): Dropout(p=0.1, inplace=False))
  )
  (intermediate): BertIntermediate(
    (dense): Linear(in_features=768, out_features=3072, bias=True))
  (output): BertOutput(
    (dense): Linear(in_features=3072, out_features=768, bias=True)
    (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
    (dropout): Dropout(p=0.1, inplace=False))
)
(8): BertLayer(
  (attention): BertAttention(
    (self): BertSelfAttention(
      (query): Linear(in_features=768, out_features=768, bias=True)
      (key): Linear(in_features=768, out_features=768, bias=True)
      (value): Linear(in_features=768, out_features=768, bias=True)
      (dropout): Dropout(p=0.1, inplace=False))
    (output): BertSelfOutput(
      (dense): Linear(in_features=768, out_features=768, bias=True)
      (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      (dropout): Dropout(p=0.1, inplace=False))
  )
  (intermediate): BertIntermediate(
    (dense): Linear(in_features=768, out_features=3072, bias=True))
  (output): BertOutput(
    (dense): Linear(in_features=3072, out_features=768, bias=True)
    (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
    (dropout): Dropout(p=0.1, inplace=False))
)
(9): BertLayer(
  (attention): BertAttention(
    (self): BertSelfAttention(
      (query): Linear(in_features=768, out_features=768, bias=True)
      (key): Linear(in_features=768, out_features=768, bias=True)
      (value): Linear(in_features=768, out_features=768, bias=True)
      (dropout): Dropout(p=0.1, inplace=False))
    (output): BertSelfOutput(
      (dense): Linear(in_features=768, out_features=768, bias=True)
      (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      (dropout): Dropout(p=0.1, inplace=Falsev
  )
  (intermediate): BertIntermediate(
```

```
          (dense): Linear(in_features=768, out_features=3072, bias=True))
        (output): BertOutput(
          (dense): Linear(in_features=3072, out_features=768, bias=True)
          (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
          (dropout): Dropout(p=0.1, inplace=False))
      )
      (10): BertLayer(
        (attention): BertAttention(
          (self): BertSelfAttention(
            (query): Linear(in_features=768, out_features=768, bias=True)
            (key): Linear(in_features=768, out_features=768, bias=True)
            (value): Linear(in_features=768, out_features=768, bias=True)
            (dropout): Dropout(p=0.1, inplace=False))
          (output): BertSelfOutput(
            (dense): Linear(in_features=768, out_features=768, bias=True)
            (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
            (dropout): Dropout(p=0.1, inplace=False))
        )
        (intermediate): BertIntermediate(
          (dense): Linear(in_features=768, out_features=3072, bias=True))
        (output): BertOutput(
          (dense): Linear(in_features=3072, out_features=768, bias=True)
          (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
          (dropout): Dropout(p=0.1, inplace=False))
      )
      (11): BertLayer(
        (attention): BertAttention(
          (self): BertSelfAttention(
            (query): Linear(in_features=768, out_features=768, bias=True)
            (key): Linear(in_features=768, out_features=768, bias=True)
            (value): Linear(in_features=768, out_features=768, bias=True)
            (dropout): Dropout(p=0.1, inplace=False))
          (output): BertSelfOutput(
            (dense): Linear(in_features=768, out_features=768, bias=True)
            (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
            (dropout): Dropout(p=0.1, inplace=False))
        )
        (intermediate): BertIntermediate(
          (dense): Linear(in_features=768, out_features=3072, bias=True))
        (output): BertOutput(
          (dense): Linear(in_features=3072, out_features=768, bias=True)
          (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
          (dropout): Dropout(p=0.1, inplace=False)))))
  (pooler): BertPooler(
    (dense): Linear(in_features=768, out_features=768, bias=True)
    (activation): Tanh()))
(dropout): Dropout(p=0.1, inplace=False)
(classifier): Linear(in_features=768, out_features=2, bias=True))
```

A.4 Demographics Comparison

Table A.3. Predicted depression ideation using CNN model on $\mathcal{P}1$ dataset

Pr.	**MN**		**NB**		**NL**		**NT**		**NS**		**ON**		**PE**		**SK**	
Sex/Age	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**
< 25	12	19	11	13	9	7	36	32	2	1	112	147	3	6	15	23
25-34	13	16	9	16	8	7	39	32	3	2	135	137	2	4	26	31
35-44	11	9	8	14	6	7	29	28	1	1	98	98	0	4	9	29
45-54	7	16	13	9	6	4	34	33	2	2	146	199	3	3	18	25
55-64	5	11	4	7	5	4	5	7	0	3	178	179	5	1	12	14
≥ 65	7	6	1	1	3	2	11	10	0	0	81	73	6	0	0	1

Table A.4. CCHS 2015-2016: Annual component for depression

Pr.	**MN**		**NB**		**NL**		**NT**		**NS**		**ON**		**PE**		**SK**	
Sex/Age	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**
< 25	1.2	0.7	0.7	0.3	0.5	0.4	0.3	0.2	0.9	0.5	5.9	4	0.3	0.2	0.9	0.6
25-34	1.2	0.6	0.5	0.2	0.5	0.2	0.3	0.2	0.9	0.5	5.7	3.3	0.3	0.1	1.2	0.7
35-44	0.9	0.6	0.5	0.4	0.6	0.3	0.2	0.1	0.8	0.5	5.7	3.2	0.3	0.1	0.7	0.5
45-54	0.9	0.7	0.7	0.5	0.6	0.3	0.1	0.2	1	0.7	6.7	4.3	0.2	0.2	0.9	0.5
55-64	1.1	0.8	0.8	0.6	0.7	0.4	0.2	0.1	1.1	0.8	6.6	4.4	0.4	0.2	0.9	0.5
≥ 65	0.6	0.5	0.5	0.4	0.6	0.3	0	0.1	0.9	0.4	4.6	2.8	0.2	0.2	0.5	0.3

Table A.5. Predicted suicide ideation using CNN model on $\mathcal{P}1$ dataset

Pr.	**AB**		**BC**		**MN**		**NB**		**NL**		**NT**		**NS**		**ON**		**PE**					
Sex/Age	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**
< 25	44	244	91	391	4	49	8	31	3	9	1	4	24	91	151	682	1	6	10	65	1	2
25-34	28	17	29	18	3	0	4	4	1	2	0	1	13	9	70	61	3	3	6	7	0	2
35-44	30	10	31	22	6	4	1	3	2	0	0	1	6	3	58	27	1	1	7	3	1	1
45-54	24	26	20	20	3	3	2	1	1	1	0	1	8	7	30	26	1	1	3	1	1	1
55-64	22	49	35	66	5	10	0	10	0	1	1	0	3	15	50	119	1	0	6	12	0	0
≥ 65	26	35	33	66	6	7	3	8	2	4	0	0	7	17	50	130	1	2	4	9	0	1

Table A.6. Canadian Community Health Survey, 2015-2016: Annual component for suicide

Pr.	**AB**		**BC**		**MN**		**NB**		**NL**		**NT**		**NS**		**ON**		**PE**					
Sex/Age	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**
< 25	2.1	1.4	1.7	1.4	1.1	0.6	0.6	0.3	0.5	0.2	0.2	0.1	0.6	0.3	4.9	3.1	0.2	0.1	0.7	0.4	0.1	0.1
25-34	1.9	1.8	1.3	1.1	0.5	0.5	0.2	0.3	0.1	0.2	0.2	0.1	0.4	0.3	2.3	1.3	0	0.1	0.5	0.5	0	0.1
35-44	1.6	0.9	1.1	0.8	0.3	0.3	0.3	0.2	0.1	0.1	0.1	0	0.2	0.1	2.5	1.4	0	0.1	0.5	0.3	0.1	0.1
45-54	1.5	1	1.3	1.1	0.6	0.2	0.2	0.2	0.2	0.1	0	0.1	0.3	0.6	3.1	2.1	0.2	0.1	0.3	0.2	0.1	0
55-64	1.3	1	1.4	0.9	0.3	0.3	0.4	0.2	0.1	0.2	0.1	0	0.6	0.3	2.4	2	0.3	0.1	0.2	0.3	0	0
≥ 65	0.6	0.2	0.6	0.6	0.3	0.1	0	0.1	0	0.2	0	0	0.2	0.2	1.2	1.2	0.1	0	0.2	0.1	0	0

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

Amayas Abboute, Yasser Boudjeriou, Gilles Entringer, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. Mining Twitter for Suicide Prevention. In *Natural Language Processing and Information Systems. NLDB 2014. Lecture Notes in Computer Science*, volume 8455, pages 250–253, Cham, 2014. Springer International Publishing. ISBN 978-3-319-07983-7.

Saeed Abdullah and Tanzeem Choudhury. Sensing Technologies for Monitoring Serious Mental Illnesses. *IEEE MultiMedia*, 25(1):61–75, 2018. doi: 10.1109/MMUL.2018.011921236.

Roberto Wellington Acuña Caicedo, José Manuel Gómez Soriano, and Héctor Andrés Melgar Sasieta. Assessment of Supervised Classifiers for the Task of Detecting Messages with Suicidal Ideation. *Heliyon*, 6(8):e04412, 2020. ISSN 2405-8440. doi: 10.1016/j.heliyon.2020.e04412.

Alejandro Adler and Martin E. P. Seligman. Using Wellbeing for Public Policy: Theory, Measurement, and Recommendations. *International Journal of Wellbeing*, 6(1):1–35, 2016. ISSN 1179-8602. doi: 10.5502/ijw.v6i1.429.

Somayyeh Aghababaei and Masoud Makrehchi. Activity-based Twitter Sampling for Content-based and User-centric Prediction Models. *Human-centric Computing and Information Sciences*, 7(1), 2017. ISSN 21921962. doi: 10.1186/s13673-016-0084-z.

Amanuel Alambo, Manas Gaur, Usha Lokala, Ugur Kursuncu, Krishnaprasad Thirunarayan, Amelie Gyrard, Amit Sheth, Randon S. Welton, and Jyotishman Pathak. Question Answering for Suicide Risk Assessment Using Reddit. In *Proceedings of the IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 468–473, 2019. doi: 10.1109/ICOSC.2019.8665525.

Hayda Almeida, Antoine Briand, and Marie-Jean Meurs. Detecting Early Risk of Depression from Social Media User-generated Content. In *Proceedings of the 8th International Conference of the CLEF Association - eRisk Shared Task (CLEF 2017)*, volume 1866, pages 1–10, Dublin, Ireland, 2017.

Salma Almouzini, Maher khemakhem, and Asem Alageel. Detecting Arabic Depressed Users from Twitter Data. *Procedia Computer Science*, 163:257–265, 2019. ISSN 1877-0509. doi: h10.1016/j.procs.2019.12.107.

Tim Althoff, Kevin Clark, and Leskovec Jure. Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics*, 4:463 – 476, 2016.

Silvio Amir, Glen Coppersmith, Paula Carvalho, Mario J. Silva, and Bryon C. Wallace. Quantifying Mental Health from Social Media with Neural User Embeddings. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 306–321, Boston, Massachusetts, 18–19 Aug 2017. PMLR.

Sairam Balani and Munmun De Choudhury. Detecting and Characterizing Mental Health Related Self-Disclosure in Social Media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15, page 1373–1378, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450331463. doi: 10.1145/2702613.2732733.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.*, 3(null):1137–1155, March 2003. ISSN 1532-4435.

Adrian Benton, Glen Coppersmith, and Mark Dredze. Ethical Research Protocols for Social Media Health Research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain, April 2017a. Association for Computational Linguistics. doi: 10.18653/v1/W17-1612.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. Multitask Learning for Mental Health Conditions with Limited Social Media Data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, pages 152–162, 2017b. ISBN 9781510838604. doi: 10.1890/06-0645.1.

Natalie Berry, Fiona Lobban, Maksim Belousov, Richard Emsley, Goran Nenadic, and Sandra Bucci. #WhyWeTweetMH: Understanding Why People Use Twitter to Discuss Mental Health Problems. *Journal of Medical Internet Research*, 19(4):107–1071, 2017. ISSN 1438-8871. doi: 10.2196/jmir.6173.

Scott R Braithwaite, Christophe Giraud-Carrier, Josh West, Michael D Barnes, and Carl Lee Hanson. Validating machine learning algorithms for twitter data against established measures of suicidality. *JMIR Mental Health*, 3(2):e21, May 2016. ISSN 2368-7959. doi: 10.2196/mental.4822.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Stephanie L Burcusa and William G Iacono. Risk for Recurrence in Depression. *Clinical psychology review*, 27(8):959–985, 2007.

Pete Burnap, Walter Colombo, and Jonathan Scourfield. Machine Classification and Analysis of Suicide-Related Communication on Twitter. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media (HT '15)*, pages 75–84, 2015.

Pete Burnap, Gualtiero Colombo, Rosie Amery, Andrei Hodorog, and Jonathan Scourfield. Multi-class Machine Classification of Suicide-related Communication on Twitter. *Online Social Networks and Media*, 2:32–44, 2017. ISSN 24686964. doi: 10.1016/j.osnem.2017.08.001.

Hugo D. Calderon-Vilca, William I. Wun-Rafael, and Roberto Miranda-Loarte. Simulation of Suicide Tendency by Using Machine Learning. In *Proceedings of the 36th International Conference of the Chilean Computer Science Society (SCCC). IEEE*, pages 1–6. IEEE, 2017. ISBN 9781538634837. doi: 10.1109/SCCC.2017.8405128.

Rafael A. Calvo, David N. Milne, M. Sazzad Hussain, and Helen Christensen. Natural Language Processing in Mental Health Applications Using Non-clinical Texts. *Natural Language Engineering*, 23(05):1–37, sep 2017. ISSN 1351-3249. doi: 10.1017/S1351324916000383.

Statistics; Canada. Canadian Community Health Survey 2015-2016: Annual Component. [Public-use microdata file]. Technical Report September, Ottawa, ON: Statistics Canada, 2017.

Patricia A. Cavazos-Rehg, Melissa J. Krauss, Shaina Sowles, Sarah Connolly, Carlos Rosas, Meghana Bharadwaj, and Laura J. Bierut. A Content Analysis of Depression-related Tweets. *Physiology & behavior*, 176(1):100–106, 2016. doi: 10.1016/j.chb.2015.08.023.A.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal Sentence Encoder, 2018.

Xin Chen, Yu Wang, Eugene Agichtein, and Fusheng Wang. A Comparative Study of Demographic Attribute Inference in Twitter. *Ninth International AAAI Conference on Web and Social Media*, pages 590–593, 2015. doi: 10.1177/2047487314541731.

Xuetong Chen, Martin Sykora, Thomas Jackson, Suzanne Elayan, and Fehmidah Munir. Tweeting Your Mental Health: an Exploration of Different Classifiers and Features with Emotional Signals in Identifying Mental Health Conditions. In *Proceedings of the 51st Hawaii International Conference on System Sciences*, pages 3320 – 3328, 2018a. ISBN 9780998133119. doi: 10.24251/HICSS.2018.421.

Xuetong Chen, Martin D Sykora, Thomas W Jackson, and Suzanne Elayan. What about Mood Swings? Identifying Depression on Twitter with Temporal Measures of Emotions. In *Proceedings of the 2018 Web Conference Companion (WWW '18 Companion)*, pages 1653–1660. Association for Computing Machinery, 2018b.

Qijin Cheng, Tim Mh Li, Chi Leung Kwok, Tingshao Zhu, and Paul Sf Yip. Assessing Suicide Risk and Emotional Distress in Chinese Social Media: A Text Mining and Machine Learning Study. *Journal of Medical Internet Research*, 19(7), 2017. ISSN 14388871. doi: 10.2196/jmir.7276.

Arman Cohan, Sydney Young, Andrew Yates, and Nazli Goharian. Triaging Content Severity in Online Mental Health Forums. *Journal of the Association for Information Science and Technology*, 68(11): 2675–2689, 2017. ISSN 23301643. doi: 10.1002/asi.23865.

Gualtiero B Colombo, Pete Burnap, Andrei Hodorog, and Jonathan Scourfield. Analysing the Connectivity and Communication of Suicidal Users on Twitter. *Computer Communications*, 73:291–300, 2016. ISSN 01403664.

Mike 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*, 16(12):1–9, 2014. ISSN 14388871. doi: 10.2196/jmir.3617.

Mike Conway, Daniel O'Connor, Daniel O'Connor, and Daniel O'Connor. Social Media, Big Data, and Mental Health: Current Advances and Ethical Implications. *Current Opinion in Psychology*, 9:77–82, jun 2016. ISSN 2352250X. doi: 10.1016/j.copsyc.2016.01.004.

Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, 2014a.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy 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*, pages 1–10, 2015a.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. CLPsych 2015 Shared Task Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology*, pages 31–39, 2015b.

Glen Coppersmith, Ryan Leary, Eric Whyne, and Tony Wood. Quantifying Suicidal Ideation via Language Usage on Social Media Glen. *Acta anaesthesiologica Scandinavica*, 49(9):1387–90, 2015c. ISSN 0001-5172. doi: 10.1111/j.1399-6576.2005.00752.x.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. Exploratory Analysis of Social Media Prior to a Suicide Attempt. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 106–117, San Diego, CA, USA, 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-0311.

Glen Coppersmith, Casey Hilland, Ophir Frieder, and Ryan Leary. Scalable Mental Health Analysis in the Clinical Whitespace via Natural Language Processing. In *Proceedings of the IEEE EMBS International Conference*, pages 393–396. IEEE, 2017.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. Natural Language Processing of Social Media as Screening for Suicide Risk. *Biomedical Informatics Insights*, 10:1178222618792860, 2018. doi: 10.1177/1178222618792860.

Glen A. Coppersmith, Craig T. Harman, and Mark H. Dredze. Measuring Post Traumatic Stress Disorder in Twitter. In *Proceedings of ICWSM*, 2014b. ISBN 9781577356578. doi: 10.1016/S1003-6326(14)63309-4.

Aron Culotta. Estimating County Health Statistics with Twitter. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*, pages 1335–1344, 2014.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F. Quatieri. A Review of Depression and Suicide Risk Assessment Using Speech Analysis. *Speech Communication*, 71:10–49, 2015. ISSN 01676393. doi: 10.1016/j.specom.2015.03.004.

Saman Daneshvar, Systems Science Program, and Computer Science. User Modeling in Social Media: Gender and Age Detection User Modeling in Social Media. Master of science, University of Ottawa, 2019.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. Social Media as a Measurement Tool of Depression in Populations. In *Proceedings of the 5th Annual ACM Web Science Conference (WebSci '13)*, pages 47–56, 2013a. ISBN 9781450318891. doi: 10.1145/2464464.2464480.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. Predicting Postpartum Changes in Emotion and Behavior via Social Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*, page 3267, 2013b. ISBN 9781450318990.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. Major Life Changes and Behavioral Markers in Social Media: Case of Childbirth. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*, pages 1431–1442, 2013c. ISBN 9781450313315. doi: 10.1145/2441776.2441937.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting Depression via Social Media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, volume 2, pages 128–137, 2013d. ISBN 9781450313315.

Munmun De Choudhury, Scott Counts, Eric J. Horvitz, and Aaron 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)*, pages 626–638, 2014. ISBN 9781450325400. doi: 10.1145/2531602.2531675.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*, pages 2098–2110, 2016. ISBN 9781450333627. doi: 10.1145/2858036.2858207.

Munmun De Choudhury, Emre Kiciman, Emre Kıcıman Georgia, and Emre Kiciman. The Language of Social Support in Social Media and its Effect on Suicidal Ideation Risk. *International AAAI Conference on Weblogs and Social Media International AAAI Conference on Weblogs and Social Media*, pages 32–41, 2017. ISSN 2162-3449.

K Denecke, P Bamidis, C Bond, E Gabarron, M Househ, A. Y.S. Lau, M A Mayer, M Merolli, and M Hansen. Ethical Issues of Social Media Usage in Healthcare. *Yearbook of medical informatics*, 10(1): 137–147, aug 2015. ISSN 23640502. doi: 10.15265/IY-2015-001.

Bart Desmet and Véronique Hoste. Online Suicide Prevention Through Optimised Text Classification. *Information Sciences*, 439-440:61–78, 2018. ISSN 00200255. doi: 10.1016/j.ins.2018.02.014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186, Minneapolis, Minnesota, 2018. Association for Computational Linguistics.

Karthik Dinakar, Henry Lieberman, Allison J B Chaney, and David M Blei. Real-time Topic Models for Crisis Counseling. *KDD DSSG Workshop*, pages 1–4, 2014.

Son Doan, Amanda Ritchart, Nicholas Perry, Juan D Chaparro, and Mike Conway. How Do You #relax When You're #stressed? A Content Analysis and Infodemiology Study of Stress-Related Tweets. *JMIR Public Health and Surveillance*, 3(2):e35, 2017. ISSN 2369-2960. doi: 10.2196/publichealth.5939.

Nawshad Farruque, Osmar Zaiane, and Randy Goebel. Augmenting Semantic Representation of Depressive Language: From Forums to Microblogs. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11908 LNAI, pages 359–375. Springer International Publishing, 2020. ISBN 9783030461324. doi: 10.1007/978-3-030-46133-1_22.

Atefeh Farzindar and Diana Inkpen. *Natural Language Processing for Social Media*, volume SYNTHESIS. Morgan &Claypool, 2 edition, 2018. ISBN 9788578110796. doi: 10.2200/S00809ED2V01Y201710HLT038.

Ethan Fast, Binbin Chen, and Michael S Bernstein. Empath: Understanding Topic Signals in Large-scale Text. In *Proceedings of the Conference on Human Factors in Computing Systems*, pages 4647–4657. Association for Computing Machinery, 2016. ISBN 9781450333627. doi: 10.1145/2858036.2858535.

Renato Miranda Filho, Jussara M. Almeida, and Gisele L. Pappa. Twitter Population Sample Bias and Its Impact on Predictive Outcomes: A Case Study on Elections. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2015)*, pages 1254–1261. Association for Computing Machinery, 2015. ISBN 9781450338547. doi: 10.1145/2808797.2809328.

Yuanbo Gao, Baobin Li, Xuefei Wangy, Jingying Wangy, Yang Zhouy, Shuotian Bai, and Tingshao Zhuy. Detecting Suicide Ideation from Sina Microblog. In *Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2017)*, pages 182–187, 2017. ISBN 9781538616451. doi: 10.1109/SMC.2017.8122599.

Manas Gaur, Amit Sheth, Ugur Kursuncu, Raminta Daniulaityte, Jyotishman Pathak, Amanuel Alambo, and Krishnaprasad Thirunarayan. "Let Me Tell You about Your Mental Health!" Contextualized Classification of Reddit Posts to DSM-5 for Web-based Intervention. In *Proceedings of the International Conference on Information and Knowledge Management*, pages 753–762, 2018. ISBN 9781450360142. doi: 10.1145/3269206.3271732.

GBD. Global Burden of Disease Study 2017 (GBD 2017) Reference Life Table. *Institute for Health Metrics and Evaluation (IHME)*, 2017. doi: 10.6069/PSS7-FB75.

Robert R German, John M Horan, Lisa M Lee, Bobby Milstein, and Carol A Pertowski. Updated Guidelines for Evaluating Public Health Surveillance Systems; Recommendations from the Guidelines Working Group. 2001. URL https://www.cdc.gov/mmwr/preview/mmwrhtml/rr5013a1.htm.

George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim J.P. P Hubbard, Richard J.B. B Dobson, and Rina Dutta. Characterisation of Mental Health Conditions in Social Media using Informed Deep Learning. *Scientific Reports*, 7:45141, mar 2017. ISSN 20452322. doi: 10.1038/srep45141.

Oliver Gruebner, Sarah R. Lowe, Martin Sykora, Ketan Shankardass, S. V. Subramanian, and Sandro Galea. A Novel Surveillance Approach for Disaster Mental Health. *PLoS One*, 12(7):e0181233, 2017a. ISSN 1932-6203. doi: 10.1371/journal.pone.0181233.

Oliver Gruebner, Martin Sykora, Sarah R. Lowe, Ketan Shankardass, Sandro Galea, and S.V. V. Subramanian. Big Data Opportunities for Social Behavioral and Mental Health Research. *Social Science & Medicine*, 189:167–169, 2017b. ISSN 02779536. doi: 10.1016/j.socscimed.2017.07.018.

Tao Gui, Qi Zhang, Liang Zhu, Xu Zhou, Minlong Peng, and Xuanjing Huang. Depression Detection on Social Media with Reinforcement Learning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11856 LNAI, pages 613–624. Springer International Publishing, 2019. ISBN 9783030323806. doi: 10.1007/978-3-030-32381-3_49.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. Detecting Depression and Mental Illness on Social Media: An Integrative Review, 2017. ISSN 23521546.

T. Hahn, A. A. Nierenberg, and S. Whitfield-Gabrieli. Predictive Analytics in Mental Health: Applications, Guidelines, Challenges and Perspectives. *Molecular Psychiatry*, 22(1):37–43, 2017.

Bibo Hao, Lin Li, Ang Li, and Tingshao Zhu. Predicting Mental Health Status on Social Media A Preliminary Study on Microblog. *15th International Conference on Human-Computer Interaction 8024*, pages 101–110, 2013.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-term Memory. *Neural computation*, 9:1735–80, 12 1997. doi: 10.1162/neco.1997.9.8.1735.

Derek Howard, Marta M. Maslej, Justin Lee, Jacob Ritchie, Geoffrey Woollard, and Leon French. Transfer Learning for Risk Classification of Social Media Posts: Model Evaluation Study. *Journal of Medical Internet Research*, 22(5), 2020. ISSN 14388871. doi: 10.2196/15371.

Xiaolei Huang, Lei Zhang, David Chiu, Tianli Liu, Xin Li, and Tingshao Zhu. Detecting Suicidal Ideation in Chinese Microblogs with Psychological Lexicons. In *Proceedings the 2014 IEEE International Conference on Ubiquitous Intelligence and Computing*, pages 844–849, 2014. ISBN 9781479976461.

Xiaolei Huang, Xin Li, Lei Zhang, Tianli Liu, David Chiu, and Tingshao Zhu. Topic Model for Identifying Suicidal Ideation in Chinese Microblog. *29th Pacific Asia Conference on Language, Information and Computation*, pages 553–562, 2015.

C. J. Hutto and Eric Gilbert. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM 2014)*, pages 216–225, 2014. ISBN 9781577356578.

Zunaira Jamil, Diana Inkpen, Prasadith Buddhitha, and Kenton White. Monitoring Tweets for Depression to Detect At-risk Users. In *Proceedings of the 4th Workshop on Computational Linguistics and Clinical Psychology*, pages 32–40, 2017.

Natasha Jaques, Sara Taylor, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. Multi-task Learning for Predicting Health, Stress, and Happiness. In *Proceedings of the NIPS Workshop on Machine Learning for Healthcare*, pages 1–5, 2016.

Jared Jashinsky, Scott H. Burton, Carl L. Hanson, Josh West, Christophe Giraud-carrier, Michael D. Barnes, and Trenton Argyle. Tracking Suicide Risk Factors Through Twitter in the US. *Crisis*, 35(1): 51–59, jan 2014. ISSN 02275910. doi: 10.1027/0227-5910/a000234.

Shaoxiong Ji, Celina Ping Yu, Sai Fu Fung, Shirui Pan, and Guodong Long. Supervised Learning for Suicidal Ideation Detection in Online User Content. *Complexity*, 2018, 2018. ISSN 10990526. doi: 10.1155/2018/6157249.

Ann John, Jane Pirkis, David Gunnell, Louis Appleby, and Jacqui Morrissey. Trends in Suicide During the COVID-19 Pandemic. *BMJ*, 371, 2020. doi: 10.1136/bmj.m4352. URL https://www.bmj.com/content/371/bmj.m4352.

M. Johnson Vioulès, B. Moulahi, J. Azè, S. Bringay, M. J. Vioules, B. Moulahi, J. Aze, S. Bringay, M. Johnson Vioulès, B. Moulahi, J. Azè, and S. Bringay. Detection of Suicide-related Posts in Twitter Data Streams. *IBM Journal of Research and Development*, 62(1):7:1 – 7:12, 2018. ISSN 21518556. doi: 10.1147/JRD.2017.2768678.

Deepali J. Joshi, Mohit Makhija, Yash Nabar, Ninad Nehete, and Manasi S. Patwardhan. Mental Health Analysis Using Deep Learning for Feature Extraction. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data (CoDS-COMAD '18)*, pages 356–359, 2018. ISBN 9781450363419. doi: 10.1145/3152494.3167990.

Sayali Shashikant Kale. *Tracking Mental Disorders Across Twitter Users.* PhD thesis, University of Mumbai, 2015.

Keumhee Kang, Chanhee Yoon, and Eun Yi Kim. Identifying Depressive Users in Twitter Using Multimodal Analysis. In *Proceedings of the 2016 International Conference on Big Data and Smart Computing (BigComp)*, pages 231–238, 2016. doi: 10.1109/BIGCOMP.2016.7425918.

Christian Karmen, Robert C. Hsiung, and Thomas Wetter. Screening Internet Forum Participants for Depression Symptoms by Assembling and Enhancing Multiple Nlp Methods. *Computer Methods and Programs in Biomedicine*, 120(1):27–36, 2015. ISSN 18727565.

Ramakanth Kavuluru, Amanda G. Williams, María Ramos-Morales, Laura Haye, Tara Holaday, and Julie Cerel. Classification of Helpful Comments on Online Suicide Watch Forums HHS Public Access. *ACM BCB*, pages 32–40, 2016. doi: 10.1145/2975167.2975170.

Jason S. Kessler. ScatterText: A Browser-based Tool for Visualizing How Corpora Differ. In *55th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations (ACL 2017)*, pages 85–90, 2017. ISBN 9781945626715. doi: 10.18653/v1/P17-4015.

Ronald C Kessler and Evelyn J Bromet. The Epidemiology of Depression Across Cultures. *Annual review of public health*, 34:119–138, 2013.

Sunghwan Mac Kim, Yufei Wang, Stephen Wan, and Cécile Paris. Data61-CSIRO systems at the CLPsych 2016 Shared Task. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 128–132, 2016.

Yoon Kim. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference (EMNLP 2014)*, pages 1746–1751, 2014. ISBN 9781937284961.

Holly Korda and Zena Itani. Harnessing Social Media for Health Promotion and Behavior Change. *Health Promotion Practice*, 14(1):15–23, 2013. ISSN 1524-8399.

Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury. Detecting Changes in Suicide Content Manifested in Social Media Following Celebrity Suicides. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media (HT '15)*, volume 2015, pages 85–94. NIH Public Access, sep 2015. ISBN 2163684814. doi: 10.1016/j.physbeh.2017.03.040.

Ugur Kursuncu, Manas Gaur, Usha Lokala, Krishnaprasad Thirunarayan, Amit Sheth, and I. Budak Arpinar. Predictive Analysis on Twitter: Techniques and Applications. *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining. Springer, Cham*, pages 67–104, 2018.

E Megan Lachmar, Andrea K Wittenborn, Katherine W Bogen, and Heather L McCauley. #MyDepressionLooksLike: Examining Public Discourse About Depression on Twitter. *JMIR Mental Health*, 4(4): e43, Oct 2017. ISSN 2368-7959. doi: 10.2196/mental.8141.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.302.

Quoc Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, volume 4, pages 2931–2939, 2014. ISBN 9781634393973.

Michael Thaul Lehrman, Cecilia Ovesdotter, Alm Rubén, and A Proaño. Detecting Distressed and Non-distressed Affect States in Short Forum Texts. In *Proceedings of the 2nd Workshop on Language in Social Media*, number Lsm, pages 9–18, 2012.

Victor Leiva, Ana Freire, Ana Maria, Freire Veiga, Victor Leiva, and Ana Freire. *Towards Suicide Prevention: Early Detection of Depression On Social Media*. PhD thesis, Pompeu Fabra University, 2017.

Diya Li, Harshita Chaudhary, and Zhe Zhang. Modeling Spatiotemporal Pattern of Depressive Symptoms Caused by COVID-19 Using Social Media Data Mining. *International Journal of Environmental Research and Public Health*, 17(14):1–23, 2020. ISSN 16604601. doi: 10.3390/ijerph17144988.

Yun Li, Tao Li, and Huan Liu. Recent Advances in Feature Selection and Its Applications. *Knowledge and Information Systems*, 53(3):551–577, 2017. ISSN 02193116.

Huijie Lin, Jia Jia, Quan Guo, Yuanyuan Xue, Qi Li, Jie Huang, Lianhong Cai, and Ling Feng. User-level Psychological Stress Detection from Social Media Using Deep Neural Network. In *Proceedings of the ACM International Conference on Multimedia (MM '14)*, pages 507–516, 2014. ISBN 9781450330633.

Huijie Lin, Jia Jia, Liqiang Nie, Guangyao Shen, and Tat-Seng Chua. What Does Social Media Say about Your Stress? In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, pages 3775 – 3781, 2016.

Huijie Lin, Jia Jia, Jiezhong Qiu, Yongfeng Zhang, Guangyao Shen, Lexing Xie, Jie Tang, Ling Feng, and Tat-seng Seng Chua. Detecting Stress Based on Social Interactions in Social Networks. *IEEE Transactions on Knowledge and Data Engineering*, 29(9):1820–1833, 2017. ISSN 10414347. doi: 10.1109/TKDE.2017.2686382.

Tong Liu, Qijin Cheng, Christopher M Homan, and Vincent M B Silenzio. Learning from Various Labeling Strategies for Suicide-related Messages on Social Media: An Experimental Study. 2017.

Xinyu Liu, Yongjun Wang, Xishuo Wang, Hui Xu, Chao Li, and Xiangjun Xin. Bi-directional Gated Recurrent Unit Neural Network Based Nonlinear Equalizer for Coherent Optical Communication System. *Opt. Express*, 29(4):5923–5933, Feb 2021. doi: 10.1364/OE.416672.

Rachel Loebach and Sasha Ayoubzadeh. Wait Times for Psychiatric Care in Ontario. *Healthcare Systems (UWOMJ)*, 86:2:48–50, 2017.

David E. Losada and Fabio Crestani. A Test Collection for Research on Depression and Language Use. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9822 LNCS, pages 28–39. Springer, Cham, 2016.

David E Losada, Fabio Crestani, and Javier Parapar. CLEF 2017 Erisk Overview: Early Risk Prediction on the Internet: Experimental Foundations. In *Proceedings of the CEUR Workshop*, volume 1866, 2017.

David E. Losada, Fabio Crestani, and Javier Parapar. Overview of eRisk 2018: Early Risk Prediction on the Internet (extended lab overview). In *Proceedings of the CEUR Workshop*, volume 2125, 2018.

David E. Losada, Fabio Crestani, and Javier Parapar. Overview of eRisk 2019 Early Risk Prediction on the Internet. *CLEF 2019. Lecture Notes in Computer Science*, 11696:49–61, 2019a. doi: 10.1007/978-3-030-28577-7.

David E. Losada, Fabio Crestani, and Javier Parapar. Overview of eRisk 2019 Early Risk Prediction on the Internet. In Fabio Crestani, Martin Braschler, Jacques Savoy, Andreas Rauber, Henning Müller, David E. Losada, Gundula Heinatz Bürki, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 340–357, Cham, 2019b. Springer International Publishing. ISBN 978-3-030-28577-7.

Kate Loveys, Patrick Crutchley, Emily Wyatt, and Glen Coppersmith. Small but Mighty: Affective Micropatterns for Quantifying Mental Health from Social Media Language. In *Proceedings of the 4th Workshop on Computational Linguistics and Clinical Psychology*, pages 85–95, 2017. doi: 10.18653/v1/w17-3110.

Meizhen Lv, Ang Li, Tianli Liu, and Tingshao Zhu. Creating a Chinese Suicide Dictionary for Identifying Suicide Risk on Social Media. *PeerJ*, 3:e1455, 2015. ISSN 2167-8359.

Minna Lyons, Nazli Deniz Aksayli, and Gayle Brewer. Mental Distress and Language Use: Linguistic Analysis of Discussion Forum Posts. *Computers in Human Behavior*, 87(May):207–211, 2018. ISSN 07475632. doi: 10.1016/j.chb.2018.05.035.

Marina Marcus, M. Taghi Yasamy, Mark van Ommeren, Dan Chisholm, and Shekhar Saxena. DEPRESSION: A Global Public Health Concern. *World Health Organization Paper on Depression*, pages 6–8, 2012.

Saman Maroufizadeh, Reza Omani-Samani, Amir Almasi-Hashiani, Payam Amini, and Mahdi Sepidarkish. The Reliability and Validity of the Patient Health Questionnaire-9 (PHQ-9) and PHQ-2 in Patients with Infertility. *Reproductive Health*, 16(1):4–11, 2019. ISSN 17424755. doi: 10.1186/s12978-019-0802-x.

Matthew Matero, Akash Idnani, Youngseo Son, Sal Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H. Andrew Schwartz. Suicide Risk Assessment with Multi-level Dual-Context Language and BERT. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3005.

Rebecca Mckee. Ethical Issues in Using Social Media for Health and Health Care Research, may 2013. ISSN 01688510.

Jonathan Mellon and Christopher Prosser. Twitter and Facebook are not representative of the general population: Political attitudes and demographics of british social media users. *Research and Politics*, 4 (3), 2017. ISSN 20531680.

Jude Mikal, Samantha Hurst, and Mike Conway. Ethical Issues in using Twitter for Population-level Depression Monitoring: A Qualitative Study. *BMC Medical Ethics*, 17(1):22, apr 2016. ISSN 1472-6939. doi: 10.1186/s12910-016-0105-5.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations Ofwords and Phrases and Their Compositionality. *Advances in Neural Information Processing Systems*, pages 1–9, 2013. ISSN 10495258.

Elham Mohammadi, Hessam Amini, and Leila Kosseim. CLaC at CLPsych 2019: Fusion of Neural Features and Predicted Class Probabilities for Suicide Risk Assessment Based on Online Posts. In *Proceedings of the 6th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2019)*, pages 34–38, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3004.

Danielle Mowery, Albert Park, Mike Conway, and Craig Bryan. Towards Automatically Classifying Depressive Symptoms from Twitter Data for Population Health. *Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 182–191, 2016.

Danielle Mowery, Craig Bryan, and Mike Conway. Feature Studies to Inform the Classification of Depressive Symptoms from Twitter Data for Population Health, 2017a. ISSN 15334406.

Danielle Mowery, Hilary Smith, Tyler Cheney, Greg Stoddard, Glen Coppersmith, Craig Bryan, and Mike Conway. Understanding Depressive Symptoms and Psychosocial Stressors on Twitter: A Corpus-based Study. *Journal of Medical Internet Research*, 19(2), 2017b. ISSN 14388871.

Danielle L Mowery, Craig Bryan, and Mike Conway. Towards Developing an Annotation Scheme for Depressive Disorder Symptoms: A Preliminary Study using Twitter Data. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology*, pages 89–98, 2015.

Serra Muderrisoglu, Oguzhan Zahmacioglu, Haluk O. Bingol, Ahmet Emre Aladag, Serra Muderrisoglu, Naz Berfu Akbas, Oguzhan Zahmacioglu, and Haluk O. Bingol. Detecting Suicidal Ideation on Forums: Proof-of-concept Study. *Journal of Medical Internet Research*, 20(6):e215, jun 2018. ISSN 14388871. doi: 10.2196/jmir.9840.

Hung Nguyen, Duc Thanh Nguyen, and Thin Nguyen. Estimating County Health Indices Using Graph Neural Networks. In *Le T. et al. (eds) Data Mining. AusDM 2019. Communications in Computer and Information Science*, volume 1127, pages 64–76. Springer Singapore, 2019. ISBN 9789811516993. doi: 10.1007/978-981-15-1699-3.

Nguyen Nguyen, Thin Nguyen, Bridianne O'Dea, Mark Larsen, Dinh Phung, Svetha Venkatesh, Helen Christensen, Nguyen Nguyen, and Thin Nguyen. Using Linguistic and Topic Analysis to Classify Subgroups of Online Depression Communities. *Multimedia Tools and Applications*, 76(8):10653–10676, apr 2017a. ISSN 13807501.

Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. Affective and Content Analysis of Online Depression Communities. *IEEE Transactions on Affective Computing*, 5(3):217–226, 2014.

Thin Nguyen, Duc Thanh Nguyen, Mark E. Larsen, Bridianne O'Dea, John Yearwood, Dinh Phung, Svetha Venkatesh, and Helen Christensen. Prediction of Population Health Indices from Social Media using Kernel-based Textual and Temporal Features. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion)*, pages 99–107. Association for Computing Machinery, 2017b. ISBN 9781450349147. doi: 10.1145/3041021.3054136.

Elaine O Nsoesie, Luisa Flor, Jared Hawkins, Adyasha Maharana, Tobi Skotnes, Fatima Marinho, and John S Brownstein. Social Media as a Sentinel for Disease Surveillance: What Does Sociodemographic Status Have to Do with It? *PLoS currents*, 8:1–26, 2016. ISSN 2157-3999.

Bridianne O'Dea, Stephen Wan, Philip J. Batterham, Alison L. Calear, Cecile Paris, and Helen Christensen. Detecting Suicidality on Twitter. *Internet Interventions*, 2(2):183–188, may 2015. ISSN 22147829. doi: 10.1016/j.invent.2015.03.005.

Jihoon Oh, Kyongsik Yun, Ji-Hyun Hwang, and Jeong-Ho Chae. Classification of Suicide Attempts through a Machine Learning Algorithm Based on Multiple Systemic Psychiatric Scales. *Frontiers in Psychiatry*, 8:192, 2017.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. Deep Learning for Depression Detection of Twitter Users. In *Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

Minsu Park, David W Mcdonald, and Meeyoung Cha. Perception Differences between the Depressed and Non-depressed Users in Twitter. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 476–485, 2013.

Michael J Paul and Mark Dredze. Discovering Health Topics in Social Media Using Topic Models. *PLoS ONE*, 9(8):e103408, aug 2014. ISSN 19326203. doi: 10.1371/journal.pone.0103408.

Michael J. Paul and Mark Dredze. Social Monitoring for Public Health. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 9(5):1–183, 2017.

Michael J Paul, Abeed Sarker, John S Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen L Smith, and Graciela Gonzalez. Social Media Mining for Public Health Monitoring and Surveillance. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 21:468–79, 2016.

Zhichao Peng, Qinghua Hu, and Jianwu Dang. Multi-kernel SVM Based Depression Recognition Using Social Media Data. *International Journal of Machine Learning and Cybernetics*, pages 1–15, jun 2017.

James W. Pennebaker, Roger J. Booth, Ryan L. Boyd, and Martha E. Francis. Linguistic Inquiry and Word Count (LIWC). *Applied Natural Language Processing*, pages 206–229, 2015. doi: 10.4018/978-1-60960-741-8.ch012.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2227–2237, New Orleans, Louisiana, 2018. ISBN 9781948087278. doi: 10.18653/v1/n18-1202.

Daniel Preot, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H Andrew Schwartz, and Lyle Ungar. The Role of Personality , Age and Gender in Tweeting about Mental Illnesses. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 21–30, 2015.

V. M. Prieto, S. Matos, M. A´lvarez, F. Cacheda, and J. L. Oliveira. Lasp-1 Regulates Podosome Function. *PloS one*, 9(1):1–10, 2014. doi: 10.1371/Citation.

Public Health Agency of Canada. *Report from the Canadian chronic disease surveillance system: Mental illness in Canada 2015*, volume 2015. Minister of Health, Ottawa, Canada, 2015. ISBN 9781100515533. doi: 10.1002/yd.20038.

Joseph H. Puyat, Arminee Kazanjian, Elliot M. Goldner, and Hubert Wong. How Often Do Individuals with Major Depression Receive Minimally Adequate Treatment? A Population-based, Data Linkage Study. *Canadian Journal of Psychiatry*, 61(7):394–404, 2016. ISSN 14970015. doi: 10.1177/0706743716640288.

Gopalkumar Rakesh. Suicide Prediction With Machine Learning. *American Journal of Psychiatry Residents' Journal*, 12(1):15–17, 2017. ISSN 2474-4662.

Diana Ramírez-Cifuentes, Ana Freire, Ricardo Baeza-Yates, Joaquim Puntí, Pilar Medina-Bravo, Diego Alejandro Velazquez, Josep Maria Gonfaus, and Jordi Gonzàlez. Detection of Suicidal Ideation on Social Media: Multimodal, Relational, and Behavioral Analysis. *Journal of Medical Internet Research*, 22(7):1–16, 2020. ISSN 14388871. doi: 10.2196/17758.

Chempaka Seri Abdul Razak, Muhammad Ameer Zulkarnain, Siti Hafizah Ab Hamid, Nor Badrul Anuar, Mohd Zalisham Jali, and Hasni Meon. *Tweep: A System Development to Detect Depression in Twitter Posts*, volume 603. Springer Singapore, 2020. ISBN 9789811500572. doi: 10.1007/978-981-15-0058-9_52.

Andrew G. Reece, Andrew J. Reagan, Katharina L.M. Lix, Peter Sheridan Dodds, Christopher M. Danforth, and Ellen J. Langer. Forecasting the Onset and Course of Mental Illness with Twitter Data. *Scientific Reports*, 7(1):1–11, 2017. ISSN 20452322. doi: 10.1038/s41598-017-12961-9.

Mark A. Reger, Ian H. Stanley, and Thomas E. Joiner. Suicide Mortality and Coronavirus Disease 2019—A Perfect Storm? *JAMA Psychiatry*, 77(11):1093–1094, 11 2020. ISSN 2168-622X. doi: 10.1001/jamapsychiatry.2020.1060.

Philip Resnik, Anderson Garron, and Rebecca Resnik. Using Topic Modeling to Improve Prediction of Neuroticism and Depression in College Students. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1348–1353, 2013.

Philip Resnik, William Armstrong, Leonardo Claudino, and Thang Nguyen. 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*, pages 54–60, 2015a. doi: 10.3115/v1/w15-1207.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-graber. Beyond LDA : Exploring Supervised Topic Modeling for Depression-Related Language in Twitter. In *Proceedings of the 52nd Workshop Computational Linguistics and Clinical Psychology*, volume 1, pages 99–107, 2015b. ISBN 9781941643167.

Hannah Ritchie and Max Roser. *Our World in Data*, 2018. URL https://ourworourworldindataldindata.org/mental-health.

Jo Robinson, Maria Rodrigues, Steve Fisher, Eleanor Bailey, and Helen Herrman. Social Media and Suicide Prevention: Findings from a Stakeholder Survey. *Shanghai Arch Psychiatry*, 27(1):27–35, 2015.

Arunima Roy, Katerina Nikolitch, Rachel McGinn, Safiya Jinah, William Klement, and Zachary A. Kaminsky. A Machine Learning Approach Predicts Future Risk to Suicidal Ideation from Social Media Data. *npj Digital Medicine*, 3(1):1–12, 2020. ISSN 23986352. doi: 10.1038/s41746-020-0287-6.

Michal Rzeszewski and Lukasz Beluch. Spatial Characteristics of Twitter Users—Toward the Understanding of Geosocial Media Production. *ISPRS International Journal of Geo-Information*, 6(8):236, 2017. ISSN 2220-9964. doi: 10.3390/ijgi6080236.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Megha Agrawal, Gregory J Park, Shrinidhi K Lakshmikanth, Sneha Jha, Martin E P Seligman, and Lyle Ungar. Characterizing Geographic Variation in Well-Being using Tweets H. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, pages 331–346, 2013a. ISBN 9783540705987.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E.P. P Seligman, and Lyle H Ungar. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, 8(9), 2013b.

H Andrew Schwartz, Johannes Eichstaedt, Margaret L Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. Towards Assessing Changes in Degree of Depression through Facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125, 2014.

Jane HK Seah and Kyong Jin Shim. Data Mining Approach to the Detection of Suicide in Social Media: A Case Study of Singapore. In *Proceedings of the 2018 IEEE International Conference on Big Data (Big Data)*, pages 5442–5444. IEEE, 2018.

Ivan Sekulić and Michael Strube. Adapting Deep Learning Methods for Mental Health Prediction on Social Media. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 322–327. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-5542.

Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat Seng Chua, and Wenwu Zhu. Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3838–3844, 2017. ISBN 9780999241103.

Leo Sher. The Impact of the COVID-19 Pandemic on Suicide Rates. *QJM: An International Journal of Medicine*, 113:707—-712, 2020. ISSN 1460-2725. doi: 10.1093/qjmed/hcaa202.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daum, and Philip Resnik. Expert , Crowdsourced , and Machine Assessment of Suicide Risk via Online Postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, 2018.

Hong-Han Shuai, Chih-Ya Shen, De-Nian Yang, Yi-Feng Lan, Wang-Chien Lee, Philip Yu, and Ming-Syan Chen. A Comprehensive Study on Social Network Mental Disorders Detection via Online Social Media Mining. *IEEE Transactions on Knowledge and Data Engineering*, 4347(c):1–1, 2018.

Lauren Sinnenberg, Alison M Buttenheim, Kevin Padrez, Christina Mancheno, Lyle Ungar, and Raina M Merchant. Twitter as a Tool for Health Research: A Systematic Review. *American Journal of Public Health*, 107(1):e1–e8, 2017. ISSN 15410048.

Ruba Skaik and Diana Inkpen. Using Twitter Social Media for Depression Detection in the Canadian Population. In *Proceedings of the 3rd Artificial Intelligence and Cloud Computing Conference (AICCC 2020)*, AICCC 2020, pages 109—-114, Kyoto, Japan, 2020a. Association for Computing Machinery. doi: 10.1145/3442536.3442553.

Ruba Skaik and Diana Inkpen. Using Social Media for Mental Health Surveillance: A Review. *ACM Computing Surveys*, 53(6), December 2020b. ISSN 0360-0300. doi: 10.1145/3422824.

Ruba Skaik and Diana Inkpen. Suicide Ideation Estimators within Canadian Provinces using Machine Learning Tools on Social Media Text. *Journal of Advances in Information Technology. Selected papers from the 13th International Conference on Machine Learning and Computing (ICMLC)*, 2020c.

Maxim Stankevich, Andrey Latyshev, Evgenia Kuminskaya, Ivan Smirnov, and Oleg Grigoriev. Depression Detection from Social Media Texts. In *Proceedings of the CEUR Workshop*, volume 2523, pages 279–289, 2019.

Maxim Stankevich, Ivan Smirnov, Natalia Kiselnikova, and Anastasia Ushakova. Depression Detection from Social Media Profiles. In *Data Analytics and Management in Data Intensive Domains.*, volume 1223 CCIS, pages 181–194. Springer International Publishing, 2020. ISBN 9783030519124. doi: 10.1007/978-3-030-51913-1_12.

Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. Detection of Depression-related Posts in Reddit Social Media Forum. *IEEE Access*, 7:44883–44893, 2019. ISSN 21693536. doi: 10.1109/ACCESS.2019.2909180.

Michael Mesfin Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. Detection of Suicide Ideation in Social Media Forums Using Deep Learning. *Algorithms*, 13(1):1–19, 2020. ISSN 19994893. doi: 10.3390/a13010007.

Georgia Tech, Munmun De Choudhury, Sanket S. Sharma, Tomaz Logar, Wouter Eekhout, and Rene Clausen René Clausen Nielsen. Gender and Cross-Cultural Differences in Social Media Disclosures of Mental Illness. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*, pages 353–369, 2017. ISBN 9781450343350. doi: 10.1145/2998181.2998220.

Robert Thorstad and Phillip Wolff. Predicting Future Mental Illness from Social Media: A Big-data Approach. *Behavior Research Methods*, 51(4):1586–1600, 2019. ISSN 15543528. doi: 10.3758/s13428-019-01235-z.

Andrew Toulis and Lukasz Golab. Social Media Mining to Understand Public Mental Health. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10494 LNCS, pages 55–70, 2017. ISBN 9783319671857.

Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. Recognizing Depression from Twitter Activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*, pages 3187–3196, 2015. ISBN 9781450331456.

Rupa Valdez and Jessica Keim-Malpass. *Ethics in Health Research Using Social Media*, pages 259–269. Springer International Publishing, Cham, 2019. ISBN 978-3-030-14714-3. doi: 10.1007/978-3-030-14714-3_13.

Kasturi Dewi Varathan and Nurhafizah Talib. Suicide Detection System Based on Twitter. In *Proceedings of the Science and Information Conference (IEEE)*, pages 785–788, 2014. ISBN 9780989319317.

Bhanu Verma, Sonam Gupta, and Lipika Goel. A Deep Learning Based Method to Discriminate Between Photorealistic Computer Generated. In *Advances in Computing and Data Sciences*, volume 19, pages 212–223. Springer Singapore, 2020. ISBN 9789811566349. doi: 10.1007/978-981-15-6634-9.

Tao Wang, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis. Detecting and Characterizing Eating-Disorder Communities on Social Media. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining (WSDM '17)*, pages 91–100, 2017a.

Yilin Wang, Jiliang Tang, Jundong Li, Baoxin Li, Yali Wan, Clayton Mellina, Neil O'Hare, Yi Chang, Neil O'Hare, and Yi Chang. Understanding and Discovering Deliberate Self-harm Content in Social Media. In *26th International World Wide Web Conference (WWW 2017)*, pages 93–102, 2017b. ISBN 9781450349130. doi: 10.1145/3038912.3052555.

Zheng Wang, Guang Yu, Xianyun Tian, Jingyun Tang, and Xiangbin Yan. A Study of Users with Suicidal Ideation on Sina Weibo. *Telemedicine and e-Health*, 24(9):702–709, 2018. ISSN 1530-5627. doi: 10.1089/tmj.2017.0189.

Zheng Wang, Guang Yu, and Xianyun Tian. Exploring Behavior of People with Suicidal Ideation in a Chinese Online Suicidal Community. *International Journal of Environmental Research and Public Health*, 16(1), 2019a. ISSN 16604601. doi: 10.3390/ijerph16010054.

Zijian Wang, Scott A. Hale, David Adelani, Przemyslaw A. Grabowicz, Timo Hartmann, Fabian Flöck, and David Jurgens. Demographic Inference and Representative Population Estimates from Multilingual Social Media Data. In *Proceedings of the World Wide Web Conference (WWW 2019)*, pages 2056–2067, New York, New York, USA, may 2019b. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313684.

Jieun Wee, Sooyeun Jang, Joonhwan Lee, and Woncheol Jang. The Influence of Depression and Personality on Social Networking. *Computers in Human Behavior*, 74:45–52, 2017. ISSN 07475632. doi: 10.1016/j.chb.2017.04.003.

Janith Weerasinghe, Kediel Morales, and Rachel Greenstadt. "Because... I was told... so much": Linguistic Indicators of Mental Health Status on Twitter. In *Proceedings on Privacy Enhancing Technologies*, volume 2019, pages 152–171, 2019. doi: 10.2478/popets-2019-0063.

Kenton White. Forecasting Canadian elections using Twitter. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9673, pages 186–191. Springer Verlag, 2016. ISBN 9783319341101. doi: 10.1007/978-3-319-34111-8_24.

Kenton White, Guichong Li, and Nathalie Japkowicz. Sampling Online Social Networks Using Coupling from the Past. In *Proceedings of the 12th IEEE International Conference on Data Mining Workshops (ICDMW 2012)*, pages 266–272. IEEE, 2012. ISBN 9780769549255. doi: 10.1109/ICDMW.2012.126.

World Health Organization. WHO. *Mental Disorders*, 2019. URL https://www.who.int/news-room/fact-sheets/detail/mental-disorders.

Sanjaya Wijeratne, Amit Sheth, Shreyansh Bhatt, Lakshika Balasuriya, Hussein S. Al-Olimat, Manas Gaur, Amir Hossein Yazdavar, and Krishnaprasad Thirunarayan. Feature Engineering for Twitter-based Applications. In *Feature Engineering for Machine Learning and Data Analytics*, page 35. Chapman and Hall, data minin edition, 2017. doi: 10.1201/9781315181080-14.

J T Wolohan, Misato Hiraga, Atreyee Mukherjee, Zeeshan Ali Sayyed, and Matthew Millard. Detecting Linguistic Traces of Depression in Topic-Restricted Text: Attending to Self-Stigmatized Depression with {NLP}. In *Proceedings of the 1st International Workshop on Language Cognition and Computational Models*, pages 11–21, 2018.

Akkapon Wongkoblap, Miguel A. Vadillo, and Vasa Curcin. A Multilevel Predictive Model for Detecting Social Network Users with Depression. In *Proceedings of the 2018 IEEE International Conference on Healthcare Informatics (ICHI 2018)*, pages 130–135. IEEE, 2018. ISBN 9781538653777. doi: 10.1109/ICHI.2018.00022.

Min Yen Wu, Chih Ya Shen, En Tzu Wang, and Arbee L.P. Chen. A Deep Architecture for Depression Detection Using Posting, Behavior, and Living Environment Data. *Journal of Intelligent Information Systems*, 54(2):225–244, 2020. ISSN 15737675. doi: 10.1007/s10844-018-0533-4.

Ming Yang, Melody Kiang, and Wei Shang. Filtering Big Data from Social Media - Building an Early Warning System for Adverse Drug Reactions. *Journal of Biomedical Informatics*, 54:230–240, 2015. ISSN 15320464. doi: 10.1016/j.jbi.2015.01.011.

Wei Yang and Lan Mu. GIS Analysis of Depression among Twitter Users. *Applied Geography*, 60:217–223, 2015. ISSN 01436228.

Andrew Yates, Arman Cohan, and Nazli Goharian. Depression and Self-Harm Risk Assessment in Online Forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, 2017.

Amir Hossein Yazdavar, Hussein S. Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. Semi-Supervised Approach to Monitoring Clinical Depressive Symptoms in Social Media. In *Proceedings of the 2017 IEEE/ACM International Conference on A2dvances in Social Networks Analysis and Mining*, pages 1191–1198, 2017. ISBN 9781450349932. doi: 10.1145/3110025.3123028.

Lei Zhang, Xiaolei Huang, Tianli Liu, Ang Li, Zhenxiang Chen, and Tingshao 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)*, volume 8944, pages 549–559, 2015.

Yunpeng Zhao, Yi Guo, Xing He, Jinhai Huo, Yonghui Wu, Xi Yang, and Jiang Bian. Assessing Mental Health Signals among Sexual and Gender Minorities using Twitter Data. In *Proceedings of the 2018 IEEE International Conference on Healthcare Informatics Workshops (ICHI-W 2018)*, pages 51–52. IEEE, 2018. ISBN 9781538667774. doi: 10.1109/ICHI-W.2018.00015.

Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. Text Classification Improved by Integrating Bidirectional Lstm with Two-dimensional Max Pooling. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, volume 2, pages 3485–3495, 2016. ISBN 9784879747020.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, Kristy Hollingshead, Ozlem Uzuner, and Kristy Hollingshead. CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts. In *Proceedings of the 6th Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-3003.