



**Arab American University**

**Faculty of Graduate Studies**

**An NLP-Based System for the Screening of Clinical Depression**

By

**Momen Hashim Abu Salameh**

Supervisor

**Dr. Mohammed A. M. Maree**

Co-Supervisor

**Dr. Mohammad Herzallah**

**This thesis was submitted in partial fulfillment of the requirements  
for the Master's degree in Data Science and Business analytics**

**May/ 2022**

**© Arab American University – 2022. All rights reserved.**

## **An NLP-Based System for the Screening of Clinical Depression**

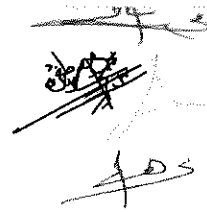
By  
**Momen Hashim Abu Salameh**

This thesis was defended successfully on 25/6/2022 and approved by:

### Committee members

1. Supervisor Name: Dr. Mohammed A. M. Maree
- B. Co-Supervisor Name: Dr. Mohammad M. Herzallah
3. Internal Examiner Name: Dr. Mujahed Eleyat
4. External Examiner Name: Dr. Rashid Jayousi

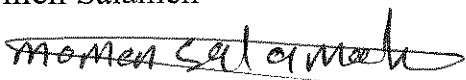
### Signature

The block contains four handwritten signatures corresponding to the committee members listed on the left. The signatures are written in black ink and are somewhat stylized.This block contains a single handwritten signature, which is more prominent and larger than the others, likely belonging to the external examiner.

## Declaration

This is to declare that the thesis entitled "An NLP-Based System for the Screening of Clinical Depression" under the supervision of Dr. Mohammed A. M. Maree and Dr. Mohammad Herzallah is my own work and does not contain any unacknowledged work or material previously published or written by another person, except where due reference is made in the text of the document.

Name: Momen Salameh

Signature: 

Date: 9/11/2022

لمن روحهم سبقتنا وتخضبت الأرض بدمائهم .. شهداءنا الأبرار  
لمن حريتهم أسرت أنعيش بكرامة ... اسرانا الأبطال  
لأم البدايات وأم النهايات ... فلسطين الحبيبة

للذين لم ييخلوا علينا بعلم او وقت المشرفون الدكتور محمد مرعي والدكتور محمد حرز الله

لمن افنى زهرة شبابه في باستيلات العدو وعاش لأجلنا مناضلا رافعاً راسنا دائما ... والدي  
العزير

لمن الجنه تحت اقدامها الداعمة والمربية ... والدتي العزيزة

للسند على صعاب الحياة أخي الدكتور محمد وعائلته

لمن فرقتنا المسافات وبقيت في أعماق الروح والقلب اختي الاستاذة مرام وعائلتها

للذي افنى عمره لنكون سويا عمي هشام وحرمة

للأب الثاني ... السد والسند عدنان مساد وعائلته

لزوجتي الحبيبة.. كلمة السر في كل المراحل والتي قاسمتني هموم الحياة ووفرت كل الظروف  
لأكون أفضل

لمن بهم تحلو الحياة وتزهو هشام وجبران فلذات كبدي

لزملائي في العمل الذين تقاسمت وإياهم مرارة الحياة لنسهر على راحة المواطنين .. زملائي  
في جهاز الامن الوقائي

أهدي هذا العمل .. والله ولي التوفيق

## Abstract

Clinical depression is a common mental disorder characterized by depressed mood and loss of interest/pleasure. Other symptoms include decreased energy, feelings of guilt or low self-worth, disturbed sleep or appetite, and poor concentration. According to the WHO, clinical depression affects over 75% of people worldwide. Usually, clinical depression screening is based on psychological evaluation through face-to-face interviews. Such conventional methods are time-consuming and error-prone. Previous research showed that it is possible to identify patients with clinical depression through the analysis of social media posts. However, analyses of social media content were neither stratified according to the various symptoms of clinical depression nor based on the exploitation of external resources of medical semantics. In addition, current screening solutions can be characterized by a loosely coupled nature, hence the low level of validity and reliability of these tools. In particular, the validity of these methods is not based on the mathematical structure of symptoms in a multidimensional construct. Accordingly, we developed a strongly coupled system that combines Natural Language Processing (NLP) and medical knowledge resources to assist healthcare professionals in screening for clinical depression; deploying a reliable and efficient tool that can passively and automatically assist in identifying subjects with clinical depression symptoms based on their social media posts. Given the dire state of mental health services and resources, the proposed system is expected to address immediate clinical relevance as it integrates the Beck Depression Inventory II (BDI-II) with social media post-analysis for both English and Arabic languages.

For the English Language, we used the Cross-Language Evaluation Forum (CLEF) eRisk 2020 dataset which is a global dataset that includes social media posts of depressed people with

different severity levels, and posts of people who have signs of Pathological gambling with signs of self-harming. Our goal in this context is to find whether social media posts reflect the presence over the Arabic social media posts, as well as severity of clinical depression symptoms if the social media posts have been written using English Language. To do this, we use multiple semantic resources in the domain of psychiatry to map the content of social media posts to their corresponding symptoms of clinical depression in the BDI-II. We utilized the outcomes of this step to train the proposed system and develop a reliable and efficient screening procedure. As the produced results indicate, the utilization of proposed pipeline and integrating it with the Bidirectional Encoder Representations from Transformers (BERT) model have resulted in 87% accuracy rate, while integrating the proposed pipeline with other classifiers, such as Logistic regression, Naïve Bays, XGBboost, Support Vector Machine (SVM) and Random Forest (RF) classifiers produced 55.8%, 50%, 47.4%, 57% and 49%, respectively.

On the other hand, for the Arabic language, we have used (Arabic Sentiment Analysis 2021 Dataset) dataset to develop an NLP-based pipeline to detect depressed subjects based on their social media posts that are written in Arabic. To evaluate the developed model, we have utilized multiple machine learning based classifiers using the Arabic dataset and achieved 81% accuracy rate using the SVM and 70% accuracy using the Logistic Regression (LR), NB and RF classifiers, respectively.

# TABLE Contents

Declaration	iii
CHAPTER One	1
Introduction	1
1.1 Background and Motivation	2
1.2 Problem Statement	4
1.3 Research Questions and Methodology	8
1.4 Thesis Organization	11
CHAPTER Two	13
Literature Review	13
2.1 Introduction	14
2.2 Theoretical Background	15
2.2.1 Sentiment Analysis in the Context of Depression Screening	19
2.2.2 Depression Screening using Machine Learning	20
2.3 Summary	29
CHAPTER Three	30
Proposed Methodology and Theoretical Framework	30
NLP Pipeline Construction Methodology	31
3.1 Dataset Introduction	32
3.2 Data Pre-processing	33
3.3 : Data Acquisition and Cleansing	34
3.4 Tokenization and Feature Extraction	35
3.4.1 Stopwords Removal	36
3.4.1 Word Stemming and Lemmatization Processes	37
3.5 Feature Extraction	38
3.4 : Machine Learning Models	44
4. Experimental Setup	49
4.1 Machine Learning <del>Parameter</del> Optimization	51
4.2 Discussion and <del>Findings</del>	52
4.3 Visualizing the <del>Results</del> Terms	67
5. Conclusions and <del>Recommendations</del>	73
References	75

## List of Figures

Figure 1: Text Mining pipeline	16
Figure 2:Sentiment Analysis workflow	25
Figure 3:Depression Severity VS posts numbers	38
Figure 4:Arabic stopwords NLTK	41
Figure 5:statistical view of the texts	43
Figure 6:Language spread over text	44
Figure 7:username posts number	44
Figure 8:Top 20 usernames vs posts	45
Figure 9:Text wordcloud	45
Figure 10: Severe Depression text wordcloud	46
Figure 11: Moderate depression wordcloud	47
Figure 12: Mild Depression wordcloud	48
Figure 13: Minimal severity wordcloud	49
Figure 14:Logistic regression	50
Figure 15:Random Forest	52
Figure 16:Maximum-margin hyperplane and margins for an SVM trained samples from two classe	53
Figure 17:BERT Model	54
Figure 18:Source: BERT (Devlin, Chang, Lee, & Toutanova, 2018),with modifications	54
Figure 19:confusion matrix	55
Figure 20: Python split function	65
Figure 21: wordnet Tokenizer	66
Figure 22Lancster lemmatizer	66
Figure 23: ISRI stemmer	67
Figure 24: Arabic nltk stopwords removal	67
Figure 25: wordnet lemmatizer	68
Figure 26: Qalsdi Lemmatizer	68
Figure 27: Logistic Regression	69
Figure 28: Random Forest over Arabic	69
Figure 29: Naive Bayes	70
Figure 30: SVM of Arabic dataset	70
Figure 31:Depression Severity percenta	71



Figure 32:Top 20 terms that have been used by the users.	71
Figure 33:Plotly top 20 used terms	72
Figure 34:20 Mild Depression most frequent words	73
Figure 35:Moderate Depression most frequent words	74
Figure 36:Minimal Depression most frequent words	75
Figure 37:Severe depression most frequent words	76

## List of Tables

Table 1: Classification of the studied MDD approaches.	30
Table 2: Logistic regression Accuracy	57
Table 3: Naive Bayes with n-gram change	57
Table 4: XGB boost Accuracy	58
Table 5: Random Forest classifier	58
Table 6: Confusion matrix of Naive model	59
Table 7: Predicted outcomes of depression severity - Naive bayes	60
Table 8: Logistic regression confusion matrix	60
Table 9: Predicted outcomes of LR	61
Table 10: Confusion Matrix of SVM	62
Table 11: SVM Prediction outcomes	62
Table 12:c Confusion Matrix of Random Forest	63
Table 13: Random Forest predicted outcomes	63
Table 14: Confusion Matrix of Gradient boosting	64
Table 15: Gradient boosting predicted outcomes	64

## **List of Abbreviations**

**NLP** – Natural Language Processing

**SA** – Sentiment Analysis

**MA**- Machine Learning

**CLEF** – Conference and Labs Evaluation Forum.

**SA**: Sentiment Analysis.

**NLP**: Natural Language Processing.

**ANNs**: Artificial Neural Network.

**FNN**: Feedforward Neural Network

**CNN**: Convolutional Neural Network.

**RNN**: Recurrent Neural Networks.

**LSTM**: Long-Short Term Memory.

**DBGRU**: Deep Bidirectional Gated Recurrent Unit.

**MLP**: Multi-Layer Perceptron.

**GloVe**: Global Vector.

**Word2Vec**: Word to Vector.

**BERT**: Bidirectional Encoder Representations from Transformers.

**SSWE**: Semantic Specific Word Embedding.

**POS**: Part-Of-Speech.

**NN**: Noun.

**ADJ**: Adjective.

**ADV**: Adverb.

**VV**: Verb.

**RL**: Reinforcement Learning

**SVM**: Support Vector Machine.

**NB**: Naïve Bayes.

**LR**: Logistic Regression.

**RF**: Random Forest.

**KNN**: K-Nearest Neighbors.

**FF:** Feature Frequency.

**TP:** True Positive.

**TN:** true negative.

**FP:** False Positive.

**FN:** False Negative.

# **CHAPTER One**

## **Introduction**

## 1.1 Background and Motivation

Clinical depression is a common mental disorder characterized by depressed mood and loss of interest/pleasure. Other symptoms include decreased energy, feelings of guilt or low self-worth, disturbed sleep or appetite, and poor concentration. In Palestine, 25% of patients are affected by clinical depression. Despite these alarming statistics, many patients with clinical depression remain undiagnosed until it is too late (Marie, SaadAdeen, & Battat, 2020). It remains possible to identify patients with clinical depression through the analysis of social media. As most social media users share their feelings and the daily lifestyle on their profiles, we can use such sources of information to potentially screen clinical depression (Sadeque, Xu, & Bethard, 2018). Mental health professionals conduct the screening for clinical depression using lengthy pen-and-paper questionnaires. The process is time-consuming, exhaustive, and inconvenient for the patient. As a result, the obtained results could be imprecise, leading to faulty screening and misdiagnosis (Eack, Greeno, & Lee, 2006; Mattsson, Olsson, Carlsson, & Johansson, 2019). To address this problem, researchers attempted to utilize social media profiles with machine learning classifiers to detect those who have an existing diagnosis of clinical depression. However, such previous efforts were focused on assembling binary models for the presence or absence of clinical depression. The search parameters were loosely coupled with the different symptoms of clinical depression, and targeting uni-languages.

In line with previous research works, we also perceive social media platforms, such as Twitter and Facebook as indispensable tools that allow people globally to share their opinions as text, video, or images. With the ever-increasing utilization of such platforms, tremendous amounts of unstructured rich text started to be generated, calling for the exploitation of various machine learning techniques for processing and understanding such textual content and to get insights

from it. As such, in our research project, we focus on the exploitation of text mining and natural language processing techniques in an attempt to derive insights from social media posts and link that to answering our main research question, namely “Will the integration of NLP pipelines and machine learning classification models assist in identifying subjects with clinical depression based on their social media posts?”.

Natural Language Processing (NLP) and the exploitation of data mining and social media analysis have recently become two hot topics in healthcare artificial intelligence. In this project, we have split our work into two main phases. Phase 1 is to do the text mining over Arabic dataset. Using the Arabic dataset, we aim to detect depressed subjects among other non-depressed subjects. On the other hand, in Phase 2, we aim to detect clinical depression using an English-based dataset (eRisk 2021) extracted from social media posts. The Arabic dataset is collected from twitter and labeled in binary format (1 is for the depressed and 0 for non-depressed users) while the English dataset is giving more details and information since the same users who wrote the posts in social media have been interviewed and answered the depression questionnaire. To accomplish this task, we used the multiclass classification models over the datasets. Based on the classifiers that we have built; we could find if the post has been written from a person who has depression and to which severity the post-writer belongs. It is important to point to the fact that by developing the proposed system, we aimed to offer a fast and accurate screening tool that replaces the need for a patient's presence in the clinic, due to the stigma of society about psychiatric patients.

## 1.2 Problem Statement

More than 300 million people suffer from depression or depression symptoms globally (AlSagri & Ykhlef, 2020). This warning statistics leads to thinking about this illness and trying to find ways that lead to early screening of depression. Normally, the procedure of screening depression is through face-face interviews with healthcare specialists and this is often tied to the stigma of the mental illness that prevents patients from going to the specialists till depression reaches a critical point, where curing depression becomes harder and takes a long duration to control the symptoms. Such conventional methods of screen depression are time-consuming and error-prone. Besides, it is important to emphasize that patients are usually reluctant to visit mental health professionals due to social stigma. To address these issues, we have proposed the utilization of technology-based solutions that can automate the screening process on the one hand, and provide more efficient, precise, and privacy-oriented methods on the other. Such solutions rely on external resources, including questionnaires and social media posts of individuals to assist in the screening process for clinical depression. And as has been recently reported by various researchers [Yalamanchili, Kota, Abbaraju, Nadella, & Alluri, 2020] , the analysis of social media posts has become among the main sources for understanding people's emotions in different situations of daily life. Classification of social media posts has been at the backbone of various systems that aim at categorizing subjects into a number of classes, including those related to clinical depression severity levels. In such multi-class classification tasks, the data is divided into several subclasses based on their polarity. Machine learning and deep learning techniques have been utilized to accomplish this task (Liu, 2012). Figure 1 shows a general architecture of a text mining pipeline that expresses the procedure of the textual content



analysis from the first stage of collecting data and ending with model preparation, which is accordingly considered as the input for training various machine learning models to predict classes that have been learned during the stage before.

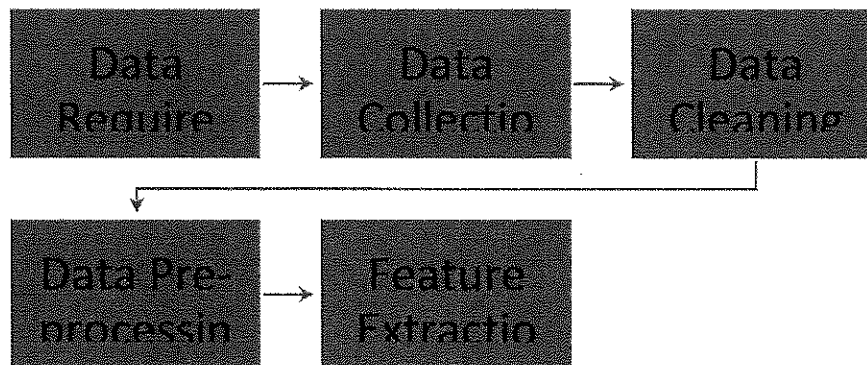


Figure 1: A General Textual Content Mining Pipeline [(Zucco, Calabrese, & Cannataro, 2017) ]

As depicted in Figure 1, first, in the Data Requirement phase, researchers specify the target of the research and how the data will help in getting the desired results. In our research project, the sample of subjects involved should have a long time presence over social media platforms, and they also should fill the depression questionnaire. Based on that, our ground-truth dataset will contain social media posts and the questionnaire results of the sample. For this reason, the data collection phase becomes crucial as it will impact the quality of the remaining processes. In our project, for English speaking subjects, we have collected a dataset that has been developed as a publicly-available dataset for depression detection purposes, specifically based on tracking posts from users who are using social media platforms, namely Reddit. In the Data Cleaning phase, we used Python programming language, in addition to the NLTK library and other manually-created functions to clean the dataset from the null values and the unnecessary characters

like the numbers and special symbols. For the Data Preprocessing phase we have used the Natural language preprocessing built-in functions in Sklearn library to do the tokenization, stemming, and lemmatization. The last phase of this pipeline is the feature extraction which prepares the dataset for the training process that machine learning models exploit for learning and further prediction purposes. In this phase, we use n-grams, Term frequency, inverse document frequency, and POS tags as part of the extracted features from the pre-processed texts.

As far as the Arabic dataset is concerned, we have obtained the dataset from Kaggle<sup>1</sup> which is a public source of multi-domain datasets, including the one that meets our research interests. The dataset was collected from Twitter for Arabic users. It is a binary-labeled dataset which means it has been labeled with 0 for non-depressed subjects and 1 for depressed ones. In the same manner as we did for pre-processing the English dataset, we used Python and a number of manually-constructed, as well as built-in functions to clean the dataset from null values and extraneous characters such as numbers and symbols during the cleaning phase. To conduct the tokenization, stemming, and lemmatization in the preprocessing step, we used the Natural language preprocessing built-in functions in the Sklearn library. Similar to the features extracted for training utilized machine learning models, we also extracted features from Arabic texts, including n-grams, stems, lemmas, and term frequencies to develop the classifiers.

Normally the analysis of social media posts produces an output that is in the form of either a 2-class result (depressed and non-depressed) or a multi-class result (mild depression, severe depression, no depression). This of course depends on the utilized training dataset, as well as the

---

<sup>1</sup> <https://www.kaggle.com/abhishek/arabic-sentiment-analysis-2021-kaust>

quality of the exploited NLP pipeline for processed texts. Conventionally, there are two techniques that are involved in such analysis tasks. These are:

- Rule based analysis: in this task, rules and dictionaries of words with labeled polarities are employed to identify the polarity of the subject based on the words prior polarity as they are defined in the used dictionary.
- Machine Learning based analysis: using this approach, machine learning algorithms are employed to extract features from sentences and detect the polarity of word mentions, and accordingly predict the class under which the subject should be classified.

Several technical challenges are faced during the above mentioned approaches. Among these challenges are:

- Being able to extract significant words for each depression class, involving a variety of NLP-based pipeline phases and evaluating the quality produced using each phase.
- Building machine learning models that have high accuracy rates for both Arabic and English languages, given that all relevant data is collected from daily posts on social media.
- Deciding on the most appropriate combination of NLP pipeline phases that should be exploited for predicting clinical depression with an aim of achieving effective and efficient machine learning models.
- Detecting the depression in Arabic language dataset. This is especially challenging due to the lack of Arabic knowledge resources.

As we discuss in the next section, as well as the remaining chapters of this thesis, we aim to address the previously highlighted challenges by building a bi-lingual clinical depression screening system that targets and analyzes social media posts written in both English and Arabic languages. In our attempt to do this, we will use multiple semantic resources in the domain of psychiatry to map the content of social media posts to their corresponding symptoms of clinical depression in the BDI-II. As will be demonstrated in the Experiments chapter, we utilized the outcomes of this step to train the proposed system and develop a reliable and efficient screening procedure. Our goal can be seen as a two-fold objective where we aim to 1) Develop an effective and efficient screening tool for assisting healthcare professionals identify clinical depression, and 2) Overcome the time-consumption, as well as error-proneness that may be produced as a result of the manual screening procedures - as is being implemented currently; targeting subjects who express their posts on social media using either Arabic or English languages.

### 1.3 Research Questions and Methodology

In this research project, we propose to develop a screening tool that integrates the Beck Depression Inventory II (BDI-II) with social media post-analysis, especially posts obtained from Reddit which have been made publicly-available as part of the CLEF eRisk 2020 dataset. In the first phase of the implemented methodology, we used a publicly-available tweets dataset labeled with depressed and non-depressed users. It is worth mentioning that due to some bias in the first obtained dataset (the bias was towards depressed subjects and specifically in terms of the words used to express depression), we have used ~~some~~ ~~data~~ that has Arabic tweets to overcome this issue as we will explain in the ~~Experiments~~ ~~section~~. As far as the English version of the dataset is concerned, we officially acquired the ~~data~~ from the CLEF eRisk 2021 organizers. As

we have pointed out earlier, the importance of exploiting this dataset lies in the fact that it saves the effort needed to manually gather posts from both depressed and non-depressed subjects on the one hand, and it is publicly-available and it enables the reproducibility of the results on the other hand. The main theme of the dataset is that it comprises posts obtained from real-world example subjects who normally express their sentiment on social media platforms, namely Reddit. To preprocess and extract features from the obtained datasets, we have developed a number of NLP-based pipelines with an aim of evaluating the impact of employing each technique among the pipeline phases on the quality of the extracted features, and how they contribute to affecting the quality of the utilized machine learning models. After constructing the pipelines, we were able to test whether social media posts reflect the presence and severity of clinical depression symptoms. Unlike conventional methods, we attempted to measure the impact of exploiting multiple semantic resources that encode clinical depression knowledge in the domain of psychiatry to map the content of social media posts to their corresponding symptoms of clinical depression in the Beck depression inventory II. We utilized the outcomes of this analysis as a training dataset for the development of a reliable and efficient screening system. This system can be used to detect the presence and severity of clinical depression symptoms passively and automatically among individuals on social media platforms.

In particular, in this research work, we attempt to answer the following research questions:

1. Is it possible to screen depression using an automatic ~~system instead~~ of manual screening procedures that are still using Depression Questionnaires?
2. Can social media posts be a relevant source for ~~mining~~ and detecting clinically-depressed subjects?

3. Is it possible to build a bi-lingual model that detects depression among both English and Arabic social media platforms?
4. Will data preprocessing pipelines affect the quality of feature extraction and accordingly impact the accuracy of the utilized machine learning models for predicting clinical depression?
5. Will dimensionality reduction techniques affect the accuracy of the system? What is the impact of changing the solver parameters of the employed machine learning models on their accuracy rates?
6. What is the impact of n-grams, TF-IDF and newer word embedding techniques on the quality of the developed models?

In our endeavor of achieving the intended research goal, this research project followed the well-recognized Design Science Research Methodology (DSRM) which was proposed by (Peffer et al., 2007). In the same manner as recommended by the DSRM methodology, the following main phases were implemented according to the chronological order described below:

- Step 1: Problem Identification and Motivation: As we described and discussed in the background of this chapter, it is reported that clinical depression is among the major diseases that calls for immediate action world-wide, and in Palestine in particular as reported by several research reports. To further understand and identify problems associated with the automatic assessment and identification of clinically-depressed subjects, we reviewed relevant literature and explored the current screening approaches, highlighting their strengths and weaknesses. We have also reviewed ~~current~~ automatic screening methods and discussed their limitations in order to ~~improve~~ ~~the~~ ~~system~~ and

specify where and how our proposed solution will contribute to existing research in this field.

- Step 2: Delineate the Objectives of a Solution: The main objectives of our research project have been set in the same manner as recommended by the DSRM research methodology. We have namely stated our main research theme, research questions in detail and how we attempted to answer each research question.
- Step 3: Design and Development of the Solution: To realize this step in our research project, we have proposed the construction of an NLP-based pipeline that comprises various phases for processing text and extracting significant features from posts of subjects involved for the preparation of both the Arabic and English versions of the obtained datasets. We have considered several tools and techniques, such as tokenization, n-gram term extracting, exploitation of medical semantic resources and various machine learning models as part of the design and development of the proposed solutions as we demonstrate in the next chapters in this thesis.
- Step 4: Evaluation: Different empirical evaluation steps were carried out as part of the evaluation phase for the proposed solution. As we detail in the Experiments chapter, we have conducted experiments using different datasets and experimental setups to evaluate each model and compare between them according to their accuracy rates.

## 1.4 Thesis Organization

The rest of this thesis is organized as follows. In Chapter 2, we introduce the related works and discuss the main characteristics that describe existing multi-class clinical depression classification techniques. Next, in Chapter 3, we provide the theoretical details and highlight

main NLP tasks that are utilized among the classification pipelines. In Chapter 4, we introduce the experimental evaluation steps and validation steps that we carried out to evaluate the utilized machine learning classifiers. Then, we present the discussion and analysis of the experiments that have been carried out using the various employed machine learning models. In Chapter 5, we present the conclusions and highlight the main future extensions to our current work.



## **Chapter Two**

### **Literature Review**

## 2.1 Introduction

Depression is a common mental disorder where as reported by the WHO, 5% of the young adults and 5.8% of old people who are above 60 years old suffers from depression<sup>2</sup>. Depression is one of the main causes for suicide globally and screening the clinical depression in early stages will be the key factor of reducing crimes globally. Based on these alarming statistics and information, building screening tools that detect depression in the early stage will help the society and the specialists in finding solutions and medications for the depression which will lead to treating the patients with the most effective manner. Major Depressive Disorder (MDD) is a mental disorder that is screened in patients who display at least five of the following symptoms of depression for at least two weeks: depressed mood, loss of interest, change in appetite or weight, guilt or worthlessness, sleep problem, psychomotor agitation or impairment, weariness or loss of energy, poor concentration, and persistent thoughts of death or suicide ideation are all symptoms of depression<sup>3</sup>. Depression and anxiety disorders affects about 1 in 15 adults in any given year, and 1 in 6 people will be affected by depression at some time in their life<sup>4</sup>. Because of the huge number of people affected by depression globally, and the stigma behind this disorder in society, especially in our Arab countries, we proposed to develop a tool that helps in screening depression online without the frequent visits to the Psychologist specialists in their clinics.

Depression screening is normally screened using Depression Questionnaires which must be filled manually by the patient who is going to the clinic. Then the doctor or the psychologists will study the output and decide if the patient has depression or not and the severity level the patient

---

<sup>2</sup> [https://www.who.int/news-room/fact-sheets/detail/depression#:~:text=Overview,world%20have%20depression%20\(1\).](https://www.who.int/news-room/fact-sheets/detail/depression#:~:text=Overview,world%20have%20depression%20(1).)

<sup>3</sup> <https://www.uptodate.com/contents/depression-in-adults-beyond-the-basics>

<sup>4</sup> <https://worldpopulationreview.com/country-rankings/depression-rates-by-country>

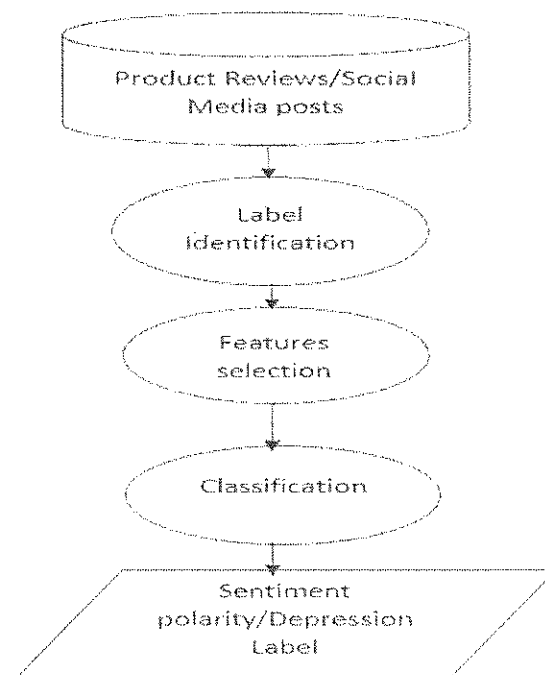
belongs to. This manual process is time-consuming and requires a lot of effort, and it can be biased according to the answers of the patient. To address these issues and save time in this context, we should think in a way that detects if the person suffers from depression or not and what's the severity of the depression if the patient has depression. As reported in the literature, one of the ways to study the behavior of a person is monitoring her/his social media posts using high performance machine learning models. In this way, patients do not need to visit the clinic physically and the model can learn from the posts which have been posted over social media for a long time by analyzing the contents since the social media account creation.

## **2.2 Theoretical Background**

Before diving into the details of our project from the technical point of view and the algorithms behind, we would like to share the theoretical information behind sentiment analysis, Machine learning methods, text mining, and natural language processing that are used to diagnose depression. Currently, researchers analyze social media content to detect the clinical depression symptoms from the posts. Therefore, we will study the content and the extent of its impact on mental illnesses, and whether it is possible to screen the medical condition through the posts on these sites. In addition to that, we will compare the results of different classifiers over the Arabic and English content of social media. The social media network phenomena generate vast volumes of valuable data that is easily accessible online. On many social networking sites, many users post photographs, videos, comments, reviews, news, and opinions, with Twitter being one of the most popular. The data collected from social media is highly unstructured, making it difficult to extract usable information from posts.

Definition 1: Sentiment Analysis (SA)

SA is defined as a computational treatment of opinions, sentiment, and subjectivity of text. The SA is one of the main Text Mining applications that are widely used nowadays. It can be used for customer satisfaction after analyzing the written text as comments (Medhat, Hassan, & Korashy, 2014). Also, it can be used in the medical field for the text mining of patient letters or posts over the internet. Figure 2 below depicts the process of SA and its output in the context of product reviews. In the context of our work, we will replace product reviews with posts obtained from social media platforms to identify clinically-depressed subjects after analyzing the content of their posts.



*Figure 2: Sentiment Analysis workflow*

## Definition 2: Machine Learning (ML)

Machine learning has been defined by IBM, one of the most important businesses working on machine learning hardware and software technologies as:

“Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.”<sup>5</sup>

As we will demonstrate in the next chapters, the utilization of various machine learning models can lead to producing varying clinical depression prediction accuracy rates. We will explore the hyper-parameters involved in this context and attempt to find out how various configurations of the hyper-parameters can change the overall quality of the proposed solution.

## Definition 3: Depression Detection

According to the huge impact of depression on a person, family, and society we need to figure out how we can know that this person has depression or not, and if yes what is the depression severity that he/she is suffering from. The main significant component of the depression diagnosis and treatment process is depression screening, because psychologists will select the treatment technique and predict outcomes based on the severity of the depression. The financial toll of the illness is significant, with coordinated rehabilitative expenses estimated at \$3.5 million per 1000 disgruntled individuals. Misery is largely underdiagnosed and undertreated, particularly in ~~essence~~ where most patients seeking help are suffering. Clinical investigations have identified ~~and~~ successful strategies for achieving successful recovery and cost reduction

---

<sup>5</sup> <https://www.ibm.com/cloud/learn/machine-learning>

over treatment. Early detection, mediation, and appropriate treatment can speed abatement, prevent relapse, and reduce the emotional and financial toll of the illness. (Halfin, 2007)

#### Definition 4: Text Mining

Text Mining can be defined as the disclosure by computer of unused, already unknown information, by consequently extricating data from distinctive written resources.

Social networks are wealthy in different kinds of content such as textual content and multimedia. The ability to apply text mining algorithms in the context of text data is important for a wide variety of applications. Social networks require text mining algorithms for many applications such as keyword search, classification, and clustering. While search and classification are well known applications for a huge number of scenarios, social networks have rich structure both in terms of text and links. Much of the work in this field uses either purely the text content or purely the linkage structure. However, many recent algorithms use a combination of linkage and content information for mining purposes. In many cases, it turns out that the use of a combination of linkage and content information provides much more effective results than a system which is based purely on either of the two. (Aggarwal & Wang, 2011)

#### Definition 5: Natural Language Processing (NLP)

The huge volume of natural language text in the connected world, though having a large content of knowledge, but it is becoming increasingly difficult to spread by a human to discover the knowledge in it, ~~specifically within~~ any given time limits. The automated NLP is aimed to do this job effectively ~~and as quickly~~, like a human does it. The challenges of NLP, progress so far made in this field ~~of~~ applications, components of NLP, and grammar of English

language—the way machine requires it. In expansion, covers the regions like probabilistic parsing, ambiguities and their resolution, information extraction, discourse analysis, NL question-answering, commonsense interfaces, commonsense thinking and reasoning, and different instruments for NLP (Chowdhary, 2020)

### **2.2.1 Sentiment Analysis in the Context of Depression Screening**

Sentiment analysis (SA) can be defined as an opinion mining process through applying natural language processing techniques over text to get the sentiment orientation of the text using rule-based and machine learning models.

Supervised and unsupervised techniques are the two main methods used in sentiment analysis research nowadays. Machine learning (ML) is often used in the supervised approach, while lexicon-based approaches are used in the unsupervised approach. The supervised technique trains a classifier or a series of classifiers using corpus data with labels (positive/negative or positive/neutral/negative). Machine learning methods such as the support vector machine (SVM), Naive Bayes (NB), Artificial Neural Networks (ANNs), and K-Nearest /neighbors (KNNs) are employed as classifiers in this context. The unsupervised or lexicon-based approach, on the other hand, defines the orientation (polarity) of each word mentioned in text as a numerical number. Some dictionaries use positive and negative numbers to distinguish between positive and negative sentimental terms, with zero indicating neutral words. The unsupervised approach, unlike the supervised approach, does not require labels.(Al Shamsi & Abdallah, 2021).

### 2.2.2 Depression Screening using Machine Learning

In a recent study by Al Asad et. al. (Al Asad, Pranto, Afreen, & Islam, 2019), the authors have applied support vector machines and Naïve Bayes classifiers on Twitter dataset which comprised 8000 tweets as labelled data for a duration of 1 year for each user. The acquired dataset has been collected by Beautiful Soup tool. After converting the dataset from JSON format to CSV they applied the pre-processing stage which is cleaning the data from the stop words and removing the unuseful symbols like (\*&%\$#). After completing the preprocessing stage, they trained the classifiers and got a 74% accuracy rate using the SVM against the Naïve Bayes classifier.

In another study by (Arora & Aeri, 2019), the researchers have used a collection of sentences obtained from Twitter with 1,850,000 tweets. The authors used multinomial Naïve Bayes and support vector regression to analyze health-related tweets for depression and anxiety detection from mixed tweets utilizing time series analysis. After getting the dataset they applied the pre-processing stage and the feature extraction to prepare the dataset as an input to the classifiers. The accuracy of the system was 78% when using Naïve Bayes, 79.7% for the SVM, and 77.17% accuracy when employing the KNN.

In the same line of research, Serra et. al. (Shrestha, Serra, & Spezzano, 2020) attempted to address the problem of detecting depressed users in online forums. The authors tracked user behaviors on the ReachOut.com online forum, which provides a safe space for young people to talk about their problems, including depression. The ~~lingua~~ ~~age~~ of user posts is analyzed in conjunction with network-based variables that ~~model~~ ~~is~~ ~~an~~ ~~extract~~ in the forum. Findings showed that network features (especially reciprocity ~~is~~ ~~a~~ ~~filtering~~ coefficient) are strong



predictors of depressed users, and that by combining them with post-linguistic features, an average precision of 0.78 has been achieved. This was different however from the utilization of other classifiers, such as the Random Forest and Random classifier with linguistic features where 0.47 for a random classifier and 0.71 for linguistic features alone were achieved, respectively.

It is worth mentioning that the utilization of NLP techniques for addressing healthcare related problems has been proposed in earlier research works as well. For instance, Avasthi et. al. and colleagues (Chakraborty, Avasthi, Kumar, & Grover, 2009) have investigated how Natural Language Processing (NLP) can be used to personalize mental health interventions. The primary goal of this study was to offer an adaptive NLP-based strategy for analyzing patient-authored text data and extracting depression symptoms using a clinically validated assessment questionnaire, the PHQ-9. To do this, the authors compared three state-of-the-art NLP algorithms with a novel word-embedding (Depression2Vec) method for extracting depression symptoms from patient-authored text. As reported by the authors, results showed that the proposed Depression2Vec embedding model performed similarly to WordNet. We would like however to point to the fact that despite the promising results achieved by the authors, WordNet and other similar lexical-based methods still suffer from missing background knowledge as reported by Maree et. al. (Maree et. al. 2018, Maree, Kmail, & Belkhatir, 2019).

With the recent advances in the development of newer word embedding techniques, (Nguyen, Nguyen, Luong, & Ngo, 2020) present two BERT fine-tuning strategies for the assumption investigation issue for Vietnamese comments, which is a strategy proposed by the BERT creators. In the context of the proposed work, the extracted features were embedded yielding expanded vectors that were used as input for the utilized models. As explained by the authors, tokens were used as input for an attached neural network,

combined with BERT, LSTM and TextCNN models. However, despite the high accuracy rates achieved, this combination - as reported by the authors - also made the training time significantly longer, recommending to avoid coupling LSTM or TextCNN with BERT. In a similar recent context, Tay et al., 2019, proposed the utilization of machine learning models for addressing mental issues and illness in an automatic manner and unlike previous manual-based models. Accordingly, the authors used machine learning models to diagnose depression through the analysis of social media posts datasets. To carry out the implementation of the proposed system prototype the authors used the Decision trees which showed improvement on the accuracy of the classification and reduced the error rate. Another machine learning model was employed by Orabi et. al. in (Orabi, Buddhitha, Orabi, & Inkpen, 2018). The authors presented a novel approach to optimize word embeddings for the classification tasks. A CNN classifier was utilized and applied on two publicly available datasets. The results produced by the used CNN model were then compared with the results of the RNN based models. As reported by the authors, For the RNN model, an accuracy rate of 82.7% was achieved, while it was 86.6% using the proposed CNN classifier.

Starting from the year 2015, research titles such as Detecting suicidality on Twitter have become more trendy, where the analysis of twitter datasets to identify users with depression symptoms that may lead to suicidality has become a main concern. As reported by O'dea et al. (O'dea et al., 2015), the research went through 14,701 suicide-related tweets. The sample processing was discovered at random among n=2000 tweets, accounting for nearly 14% of all tweets. Then the researchers labeled the dataset into 2 categories. Sixty-six percent of the tweets were "potentially concerning," while the rest were classified as "safe to be concerned".

classifiers identified 80% of the “Strongly concern” using SVM and LR with filters and without filters.

As we can notice in the literature, most of the focus has been on analyzing English-based social media posts. This indeed has been the case for several years ago until we started to find some new research works that pay attention to the importance of analyzing Arabic-based social media posts as well. However, Building a corpus for the Arabic region is a key in MENA countries especially that this region has multi dialects which may affect the performance of the models (Zaghouani, 2018). For instance, in 2019, a recent research was published by Arabic researchers to detect depression from social media. To train their machine learning algorithms, the researchers used data from Arab Gulf countries. The dataset was binary labeled data with labels (depressed , non-depressed ). They have used machine learning algorithms like (Random Forest, Naïve Bayes, AdaBoostM1, and Liblinear) to detect the depressed tweets. Using the listed algorithms the researchers found the optimal accuracy is coming as an output of applying a bilinear classifier which was 0.87 accurate (Almouzini & Alageel, 2019).

Table 1 summarizes the most relevant research works that we have reviewed prior and during the development of our proposed thesis project. As described in Table 1, the best accuracy over the models is obtained based on the exploitation of BERT which will be used in our proposed system, specifically while processing and embedding the English dataset.

Table 1 Classification of the Studied MDD approaches.

Ref	Data Size	Platform	Features	Approach		Performance				Main Characteristics
						Accuracy	Precision	Recall	F1	
O'dea et al., 2015	14,701	Twitter	TF-IDF	Machine Learning Approach	SVM And LR	0.67	0.88	0.64	0.74	The twitter data is unable to provide the characteristics like age and gender so the generalization of the study will be limited.
Almouzni &	6122	Twitter	Bag of ngrams	LibLinear		0.87	0.86	0.87	0.87	The researcher

Alage el, 2019			And negation handling						has done manual labeling over the whole dataset which can be done automatic ally by linking the users who wrote the twitter with the depressio n questionn aire result.
(Priya, Garg,	348 users	Goog le	N\A	SVM	0.803	0.82	0.7 1	0. 76	The preproces

& Tigga, 2020		forms							sing stages are not clearly mentione d during the research.
Arora & Aeri, 2019	1,850, 000	Twitt er	TF-IDF	Naïve	78%				The proposed algorithm s lead to good result over the depressio n severity detection from tweets.
				SVM	79.7%				
				KNN	77.17 %				
Al Asad,	8000	Twitt er	TF-IDF	SVM	N/A	N/A	N/ A	N/ A	The researcher

Pranto, Afreen, & Islam, 2019				Naïve Bayes	0.74	100%	0.60	NA	mentioned the SMV during the research but without clearly mentioning the performance matrix of the model.
Nguyen, Nguyen, Luong, & Ngo, 2020	50,760		BERT	SVM	89.23	92.52	90.		Researchers achieved good results.
				XGBoost	88.76	90.58	84		
				FastText+TextCNN	67.9	89.1	77.		
				FastText+LS	88.5	89.7	1		
				TM	89.2	91.7	89.		Limitation: Heavy

				FastText+R CNN			90. 4		models will cost high specificati ons hardware tools.
				Glove+Text CNN Glove+LST M Glove +RCNN	69.7 88.7 85.5	87.7 91.8 85.8	77. 7 89. 8 90. 7		
				BERT-base BERT- LSTM BERT- TextCNN BERT-RCNN	88.13 89.78 88.85 88.76	94.02 92.08 93.14 93.68	90. 9 90. 91 90. 94 91. 15		
Orabi, Buddh itha,	1,145 users	Twitt er	Word_em beddig	CNN	88	87	87. 4	87	The researcher s should compare between the RNN



, 2018									and the CNN to compare the results of the models.
--------	--	--	--	--	--	--	--	--	--

### 2.3 Summary

In this chapter, we reviewed a number of research works that are related to depression detection using machine learning models. In order to identify the main features that characterize each of the reviewed systems, we have considered a wide range of research works that targeted both English and Arabic social media datasets. During the literature review task, our main focus was on understanding the main techniques and approaches that have been implemented to address the issue of automatic screening of clinical depression. Based on the exploration and findings of this phase, we were able to highlight the main strengths, as well as weakness of existing approaches on the one hand, and inspire the backbone and core NLP pipeline phases that can be integrated and combined with word embedding techniques for the analysis and screening of clinical depression.

## **Chapter Three**

### **Proposed Methodology and Theoretical Framework**

## NLP Pipeline Construction Methodology

Before we proceed with introducing the proposed NLP pipeline's construction methodology, we would like to emphasize the importance of utilizing a publicly-available and well-recognized dataset in the field of clinical depression screen. This was among the first and most important tasks in our research project. In this context, and after reviewing the literature, we were able to gather datasets that target both Arabic and English languages, with some bias in the Arabic dataset which causes some results to skew as we will describe in the Experiments chapter. Concerning the English version of the dataset, social media posts about real-world subjects were provided globally under the title eRisk dataset. The dataset addresses three main societal problems which are: signs of pathological gambling, depression detection and signs of eating disorders. In our project, we went through the depression detection task which has been named in eRisk T2-2020. In this task the dataset contains the social media content of sample subjects who already filled the manual depression questionnaire. The dataset is labeled with four categories based on the severity of the depression starting with minimal depression, moderate, mild, and severe.


The dataset gathers the textual posts of the users who were published on Reddit social media platform<sup>6</sup>. The total number of users were 70 users, and the number of posts were 35000 posts which can be considered as a good dataset to initiate and further train the proposed screening system.

---

<sup>6</sup> <https://reddit.com/>

### 3.1 Dataset Introduction

eRisk is owned by CLEF whose main goal is to look at issues like assessment approaches, execution measures, and other issues that come up when building testbeds for early chance detection and prediction. The prescient apparatuses created beneath eRisk's shared tasks may be possibly valuable in several ranges, especially those related to health and safety. For illustration, caution alarms can be sent when an individual starts broadcasting self-destructive contemplations on social media. eRisk tries to instigate interdisciplinary investigation (e.g., related to data recovery, machine learning, psychology, and computational etymology), and the progresses created under this challenge would be possibly appropriate to bolster several socially important issues. In other words, given a stream of information (e.g., real-time Social Media entries), cautions should be lifted if sufficient proof of a particular risk has been established.

In our proposed solution, and to experimentally validate our proposal, we used task 2 of the eRisk datasets, which is related to diagnosing depression from social media profiles and manually filling out depression questionnaires. This activity comprises measuring the intensity of a few sadness-related negative consequences in a natural way. A survey with 21 questions about various emotions and well-being (e.g., sorrow, pessimism, and weariness) is offered for this purpose. Each address contains four to seven possible responses, each of which corresponds to a different level of seriousness (or relevance) of the side effect or behavior. A sample of subjects with their answers to the survey and their works at Reddit was given. To benchmark the distinctive approaches, a modern set of subjects and their writings  which each group has got to anticipate their answers.

### 3.2 Data Pre-processing

As we mentioned in the previous section, we have two datasets, where one of them has the user posts over reddit, and the dataset has a subject ID. Also, we do have another dataset that has the results of the Depression questionnaire of the users which has an attribute named subjectID. We merge these 2 datasets using the subjectID to have them in one dataset where the label of the dataset is the result of the manual Depression Questionnaire.

After counting the number of the posts per Depression severity, we found that the number of posts that has Mild Label is 11858 posts, and the number of the posts that has moderate label are 10476 posts, the severe posts were 7610, and the number of the minimal posts were 5618 posts as depicted in Figure 3.

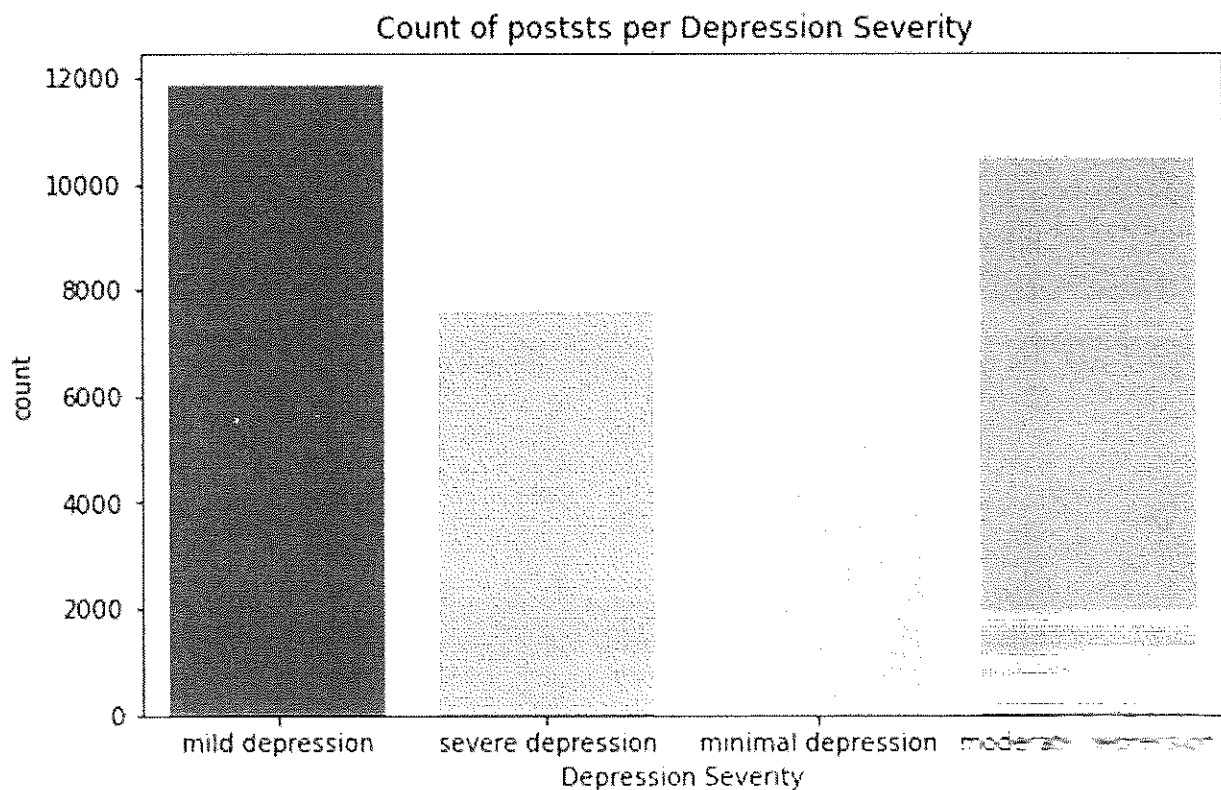


Figure 3: Depression Severity VS posts numbers

As a conventional task in NLP, the dataset must be cleaned and pre-processed to prepare it to be the training dataset to our machine learning model, especially that the machine learning model has different classifiers that all aim to give the best accuracy rates. We removed features that do not affect the system's correctness, such as the answer marks for each questionnaire question, because the most essential variable in this is the questionnaire's final evaluation. Also, we have deleted the title of the posts of the social media content and the info since all the posts were collected from a single source, that is Reddit.

For the textual content of posts, which is the most important variable in our analysis, we have done some cleaning and preprocessing tasks. For instance, before we count the number of words (term frequency) in each text, we removed the stop words which are consuming a lot from the text and it's not affecting the content of the posts. Then, we have removed the numbers and the null values from the text since the machine learning model input must have useful data to have the best accuracy.

### **3.3 : Data Acquisition and Cleansing**

Analysis of social media posts generated by real-world subjects in the context of screening depression can be seen as a data mining task that aims to calculate the sentiment orientation of a person's personality through the help of natural language processing. For analyzing the content, computational linguistics is utilized to infer and analyze mental information of Web, social media, and related references. Sentiment analysis in this regard has been utilized in numerous domains, such as healthcare and medical practices. In reality, there is a gigantic volume of healthcare data online, such as individual blogs, social media, and on the websites approximately

restorative issues rating that are not obtained methodically. Opinion analysis provides numerous benefits such as utilizing therapeutic data to realize the leading result to extend healthcare quality (Abualigah, Alfar, Shehab, & Hussein, 2020). For the evaluation process, we have used a well-known dataset (CLEF eRisk) that is used in several mental problems. This is for English speaking subjects. For the Arabic dataset, we have used public dataset which has been downloaded from Kaggle<sup>7</sup>. The English dataset has 35423 posts which have been collected from Reddit and for the Arabic 20,000 posts that have been collected from Twitter for Arab users. 50% of the posts that have been collected from the Arabic dataset are labeled as negative posts and the others are positive. The first task after the data acquisition is to clean the raw data as it normally contains some special characters, including hashtags, consecutive white spaces, URLs, and unnecessary symbols. In addition, there is a set of emoticons that we cleaned using a predefined set of icon representations. After this step, we proceed to the second phase of data processing, that is tokenization and feature extraction.

### **3.4 Tokenization and Feature Extraction**

After the stage of Data Acquisition and Cleansing, it's the time to start with the first stage of NLP Pipeline, which is the tokenization which can be affecting the final accuracy of the models that have been built based on the machine learning algorithms (Deshpande & Rao, 2017). At this point, the text will be converted into tokens, which are specific words. We discovered that the bigram gives the best accuracy on the outcome after using the ngram tokenizer. Especially when our English dataset offers multilabel depression severity classification. We tested numerous tokenization approaches on the Arabic dataset and discovered that the tweet tokenizer (Ciccone

---

<sup>7</sup> [www.kaggle.com](http://www.kaggle.com)

et al., 2018), which was created for Twitter datasets, provided the best accuracy. In terms of computational cost, using n-grams with n greater than 3 was very time consuming. In an earlier study in 2004, Pang et. al. have shown that unigram features are more significant than bigrams when performing emotional classification for movie reviews (Pang & Lee, 2004). On the other hand, other studies showed that coupling bigrams and trigrams based on extrinsic semantic resources provided better results than unigrams alone (Maree, 2021). Considering these conclusions, it was crucial to experimentally evaluate the utilization of n-gram tokenization as part of the large-scale SA process.

### **3.4.1 Stopwords Removal**

Normally, text in social media has words that do not have useful information especially for training the machine learning models. One of the main important phases of the preprocessing is the stop words removal which aim to reduce the number of words that do not affect the meaning of the text and consume the time and the effort of the program in analyzing data that will not impact the output of the model (Kaur & Buttar, 2018).

The stopwords can be divided into two main categories: 1) the general stop words like (in, an, the, for, on, of, to,...etc) these stopwords are not related to any domain of work, and can be considered as general stop words that should be deleted from the raw text to prepare the text to be an input to the machine learning model. 2) domain-specific stopwords are those words that show up to be of small importance in determining the meaning of a given sentence in a specific domain. Usually, researchers use the TF-IDF to find all the domain stopwords.

For the Arabic dataset, one of the most important steps is to apply the stopwords filter over the raw text since Arabic language has a significant number of stopwords that will not affect the



system, but they can be costly for the machine (Alajmi, Saad, & Darwish, 2012). In the figure below, we list some of the Arabic stop words that we deleted from the Arabic dataset. In the project, we used the NLTK library in Python to delete all stop words.

```
In [20]: stopwords=nltk.corpus.stopwords.words("arabic")
print(stopwords)
stopwords.extend(['', 'ا', 'آ', 'ؤ', 'ئ'])
def remove_stopwords(txt_tokenized):
    tweet_clean=[word for word in txt_tokenized if word not in stopwords]
    return tweet_clean
df2["no_stopwords"]=df2["Lanc_stemmer"].apply(lambda x: remove_stopwords(x))
df2
```

[illegible]

Figure 4: Deleted Arabic Stop Words using the NLTK Library

### 3.4.1 Word Stemming and Lemmatization Processes

In this phase, we use the raw text as input to stem and lemmatize the words. As reported by Hull (Hull, 1996), stemming is mainly concerned with the removal of derivational affixes in the hope of achieving a common base form for a given word (Porter, 2001). The objective of this process is to decrease inflectional shapes and obtain a common base form that expresses words in opinion sentences. After applying this stage, the text dimensionality will be reduced for the ML classifiers. This reduction in word dimensions makes a difference also to accurately decide the weights of the words and their significance within the content. In our project, we have used several techniques from the NLTK to test the best stemmer to be used in our systems .

Compared to stemming, lemmatization is another common morphological examination handle that has been broadly utilized for opinion examination purposes. It is basically utilized to remove inflectional endings to return the base or lexicon form of a word, which is known as the lemma. One of the most important points of interest of lemmatization is that a returned lemma can be normalized to get equivalent words, as well as other semantically-related terms from outward semantic assets such as WordNet and Yago (Plisson, Lavrac, & Mladenic, 2004).

### 3.5 Feature Extraction

For text mining and natural language processing, we have used ngrams. Ngrams consist of sets of co-occurring words within a specific threshold. Most of the work that has been done over the T2 task and the depression detection and the disorder of eating have used the ngrams. In the implementation phase, a tf-idf vectorization was done from the unigrams, bigrams, and trigrams. The scikit-learn Python library's TfidfVectorizer was used in this stage, with a stop-words list and the expulsion of n-grams that appeared in less than 20 records. The substance of a record was characterized by the concatenation of all the compositions of a client from all the chunks, within the preparing phase. Trigrams did not improve the outcomes and thus they were not utilized in the employed models.

In particular, we utilized the well-known TF-IDF weighting procedure with the combination of n-gram tokens and semantic highlights extricated from WordNet. As such, a sentiment sentence is assigned a weight based on Eq. 1. The unit we use to represent each sentence is D. We decided to use the variable D to denote sentences in the dataset to maintain consistency with the general terms used to define the TF-IDF equation.

$$TF-IDF(t) = (1 + \log_{10} TF(t)) * \log_{10}(N/DF(t))$$

Where,

- N: is the number of sentiment sentences in the corpus.
- DF: represents the number of sentences that contain the term t.
- TF: is the number of occurrences of term t in sentence D.

It is important to point out here that the TF-IDF feature weighting technique is applied on composite features extracted from pre-processed text of sentiment sentences.

To understand the dataset more, we have built the statistical description of the dataset, to see the number of texts per user and to show the statistical features like mean, median, mode, and standard deviation. For the employed models, these features appeared to supply significant outcomes at the training phase.

```
count      70.000000
mean      475.771429
std       423.641056
min        16.000000
25%        77.250000
50%       334.500000
75%       984.750000
max      1191.000000
Name: ID, dtype: float64
```

Figure 5: Statistical Overview of the Texts

In the context of the proposed solution, there was a need to detect the language of the text prior to starting the pre-processing and analysis phases. In other words, it's important to make sure that the text for all the users were written in the same language to apply the same methods over the text and make sure that all the models input are the same for each user, the figure below is the language spread over the text. We noticed that most of the text was written in English Language.

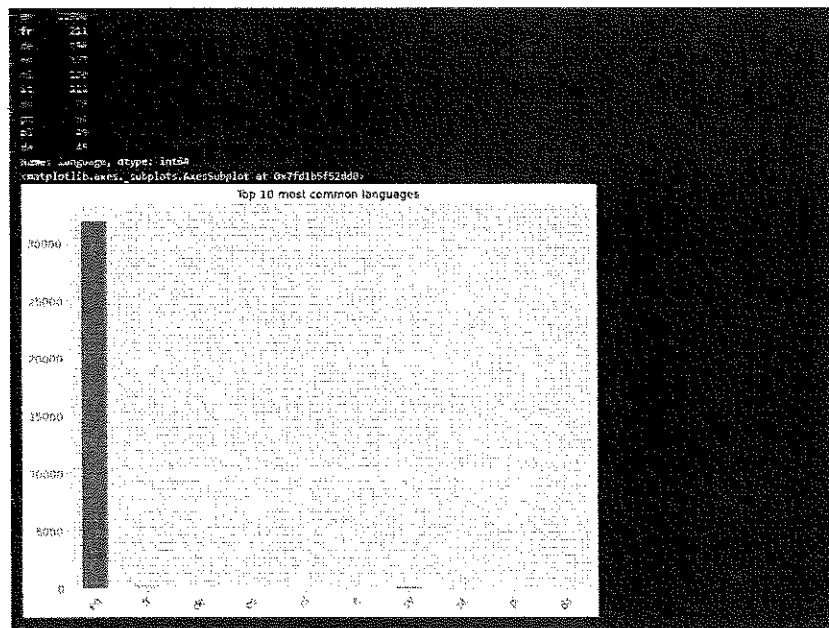


Figure 6: Language distribution over text

It's also important to know if the users have the same number of posts compared with other users. We have noticed that even the number of posts per user was different from user to another which is logical to occur in the dataset since the dataset is collecting the posts in specific timeline, the figure below is giving the username and the number of posts.

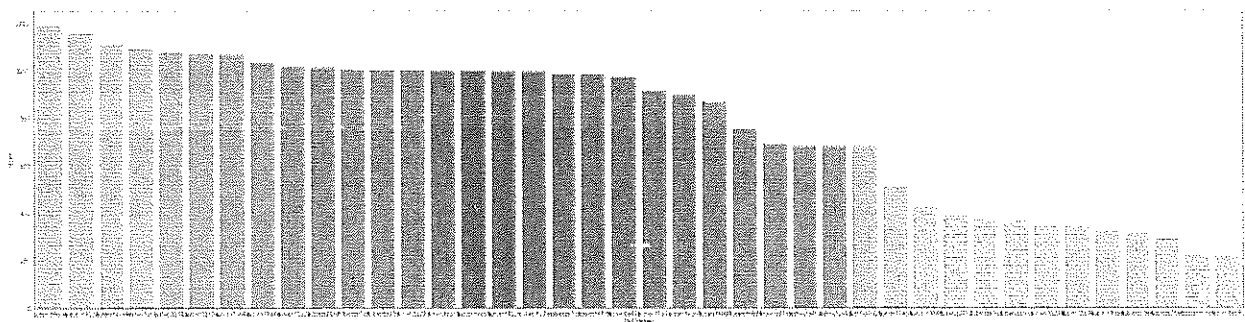


Figure 7:username posts number

Since the number of users are 70 and the number of posts is close in some cases , we decided to show ~~the number of posts~~ with the number of the posts. the figure below is showing the top 20 ~~users~~ ~~users~~

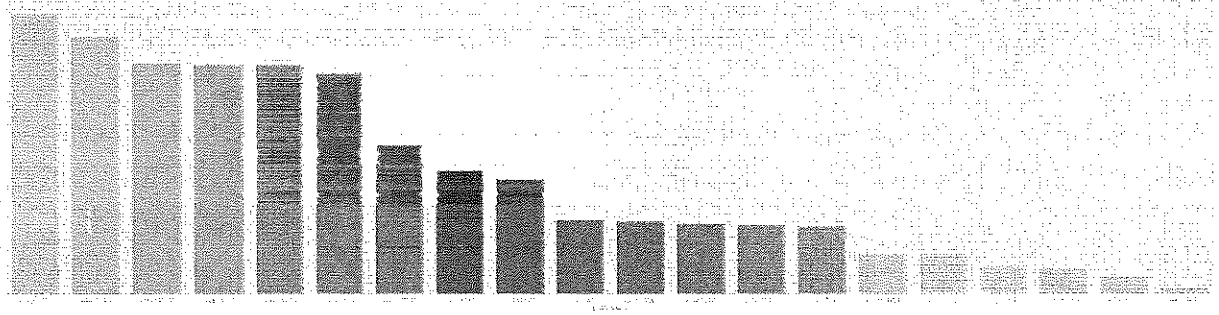


Figure 8: Top 20 usernames vs posts

For the text mining purposes, we have built the word cloud which will show an indication about the most frequent words that have been used in the text of the usernames, the figure below is showing the word cloud.



Figure 9: Text Wordcloud

We have performed wordcloud analysis for each severity of the depression, where we start with the severe depression that we labeled based on the analysis that is acquired from manually-filled depression screening questionnaire.

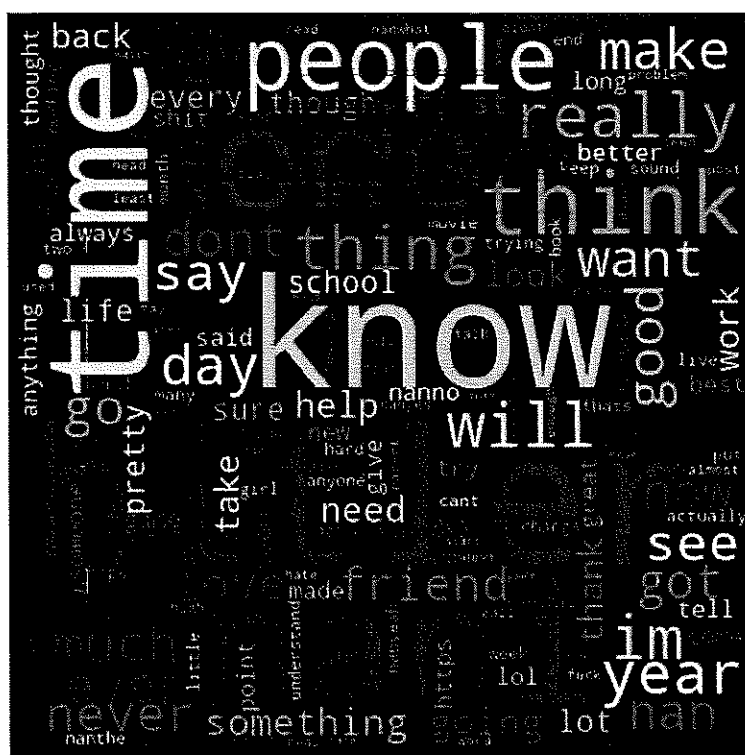


Figure 10: Severe Depression text Wordcloud

For the moderate severity, the wordcoud is shown in Figure 11 below :



Figure 1. Second

For the Mild depression, Figure 12 shows the ~~the~~ extracted from mild-severity subject posts.

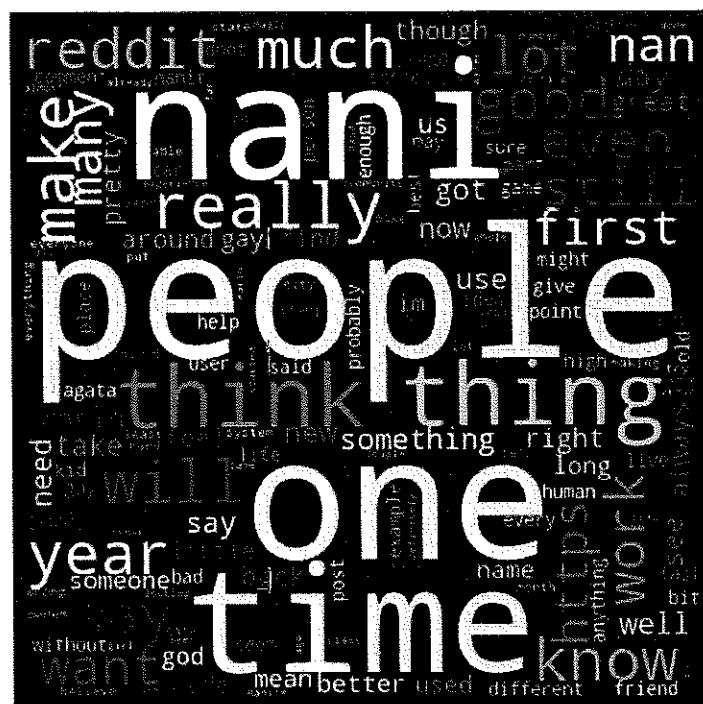


Figure 12: Mild Depression Wordcloud

For the minimal severity, Figure 13 depicts the wordcloud of minimal severity posts analysis.



Figure 13: Minimal Sequence

### 3.4 : Machine Learning Models

In the machine learning phases, we prepared the data by splitting the dataset into training and testing which were 0.8 for the training and 0.2 for the testing. The models that have been used were six classifiers which are: Logistic regression, Naïve Bayes, Random Forest, support vector machine, XGB classifier, and BERT.

Logistic regression is a statistical method used to predict the outcome for given independent variables. It predicts the probability of occurrences of events by fitting the data inputs to a Logit function where the natural log of the odds that Y equals one of the categories. The Logistic regression algorithm does not require linear relations between inputs and output variables. This will be based on applying a nonlinear log transformation to the odd ratio.

$$\text{Logistic regression} = \frac{1}{1+e^{-x}} \quad (1)$$

Input features

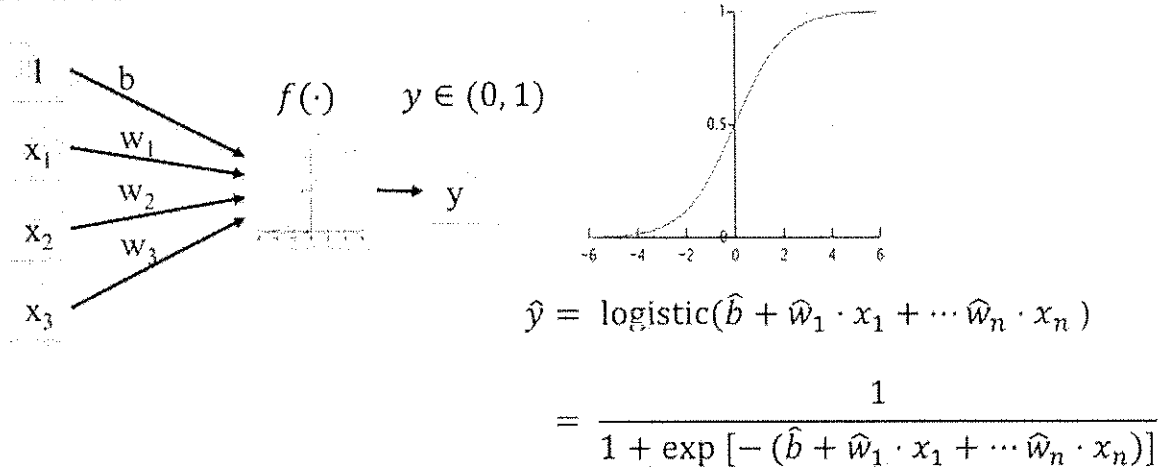


Figure 14: Logistic regression

Naïve Bayes on the other hand is a probabilistic classifier that is based on the Bayes theorem with naïve independence assumption between the independent variables. The classifier



requires several parameters linear in the number of the variables. It gives the instance a class label. The class label is chosen from a finite set, and the problem instance is represented as a vector of feature values. There is no one algorithm for training such a classifier; rather, it is a collection of algorithms based on the same premise. Class variables are present in all naive Bayes classifiers. The equation below shows the naïve bayes algorithms from the mathematical point of view.

$$P(features) = \frac{p(class)*p(class)}{p(features)} \quad (2)$$

Where  $p(class)$  it is the conditional probability of a class in corpus,  $p(class)$  is the conditional probability of a feature related to a class and  $p(features)$  is the Likelihood of a feature.

In our system, we employed the multinomial naïve bayes that implements the naive Bayes algorithm for multinomially. It is one of the two classic Naive Bayes variants used in text classification and works with scattered data. The mathematical representation of the algorithm is shown in the equation below.

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

Where  $\theta_y$  is parametrized vector for each class of  $y$ . where  $n$  is the number of features. And  $N_{yi}$  and  $N_y$  are represent below.

$$N_{yi} = \sum_{x \in T} x_i$$

$$N_y = \sum_{i=1}^n N_{yi}$$

Which means the number of times feature  $I$  appears in a sample of class  $y$  in the training set  $T$ . The smoothing priors  $\alpha \geq 0$  accounts for features do not present in the learning samples and prevents zero probabilities in further computations. Setting  $\alpha=1$  is called Laplace smoothing, while  $\alpha < 1$  is called Lidstone smoothing.

The third classifier we employ is the Random Forest, which is a categorization method that works by training a large number of decision trees. For activities requiring classification, it outputs the class that is the mode of the classes of the individual trees.

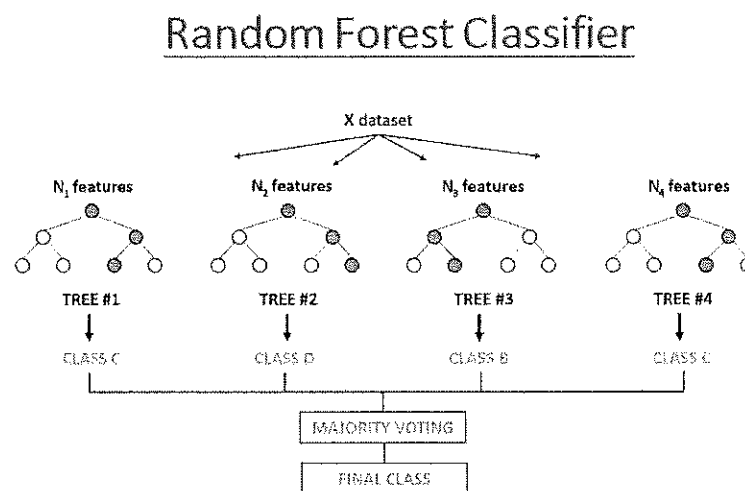


Figure 15: Random Forest Classifier

In a tall or infinite-dimensional space, a support vector machine creates a hyperplane or series of hyperplanes that can be used for classification., relapse, or other assignments like sentiment

detection. A great partition is accomplished by the hyperplane that has the biggest remove to the closest training-data point of any course (so-called utilitarian edge), since in common the bigger the edge, the lower the generalization blunder of the classifier (Agarwal, Xie, Vovsha, Rambow, & Passonneau, 2011; Alam & Yao, 2019; Hemalatha, Varma, & Govardhan, 2013).

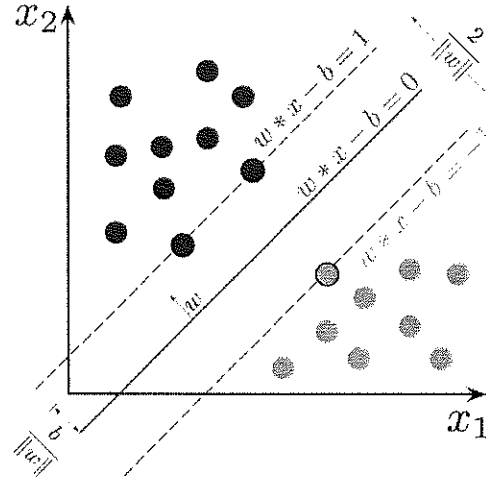


Figure 16: Maximum-margin hyperplane and margins for an SVM trained with samples from two classes<sup>8</sup>

This method separates these two sets of nonlinear data by finding the best surface separating them. With the aim of maximizing the margin between the two classes, as maximizing the margin reduces indecisive decisions. The hyperplane which separates the two classes (positive and negative classes) calculated as follow:

$$(\vec{x} \cdot \vec{y}) + b = \sum_i y_i a_i (\vec{x} \cdot \vec{y}) + b = 0 \quad (2)$$

Where  $\vec{x}$  is n-dimensional input vector and  $\vec{y}$  is output value (positive, negative).

$(\vec{x}_{in}, \vec{y}_l)$  together form a training dataset and  $a_i$  is a learning multiplayer

$$\vec{x} = (x_{i1}, x_{i2}, x_{i3} \dots, x_{in}) \quad (3)$$

$$\vec{y} = (y_1, y_2, y_3 \dots, y_i) \quad (4)$$

Using eq.1 to construct hyperplane, its represented by vector  $\vec{w}$ .

$$\vec{w} = (w_1, w_2, w_3 \dots, w_n) \quad (5)$$

Prediction classes determined at testing phase by following equations:

$$\text{Positive if } \vec{w} \cdot \vec{x} + b \geq 0 \quad (5)$$

$$\text{Negative if } \vec{w} \cdot \vec{x} + b < 0 \quad (6)$$

<sup>8</sup> [https://en.wikipedia.org/wiki/Support-vector\\_machine](https://en.wikipedia.org/wiki/Support-vector_machine)

We also employ XGB boost, which is an optimized conveyed gradient boosting library planned to be profoundly productive, adaptable, and versatile. It actualizes machine learning calculations beneath the Angle Boosting system. XGBoost gives a parallel tree boosting (also known as GBDT, GBM) that illuminates numerous information science issues in a quick and precise way.

Additionally, BERT was utilized for the purpose of bidirectional preparation of language Transformers, which are different from past endeavors which looked at a content arrangement either from cleared out to right or combined left-to-right and right-to-left preparing. Figure 17 below shows the BERT model.

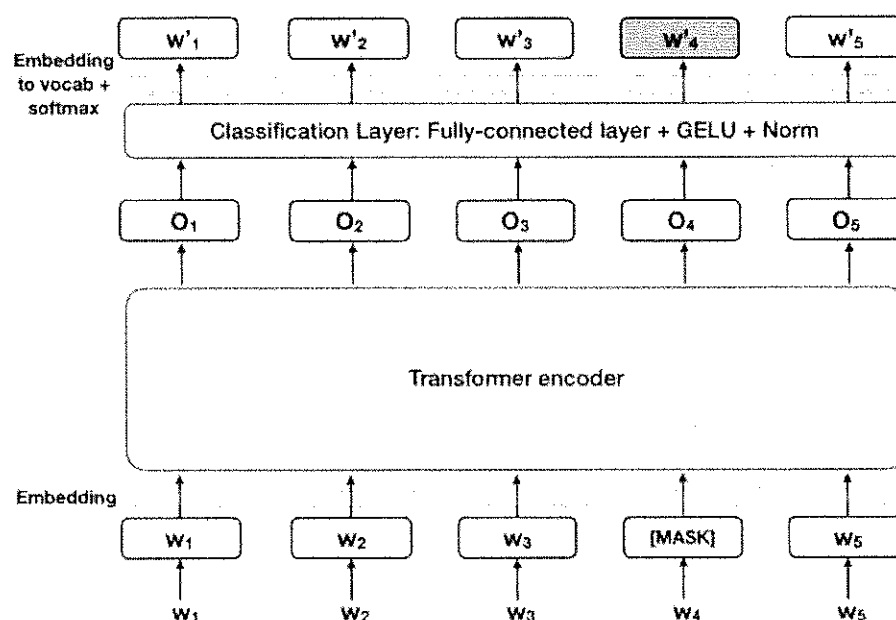


Figure 17. BERT Model

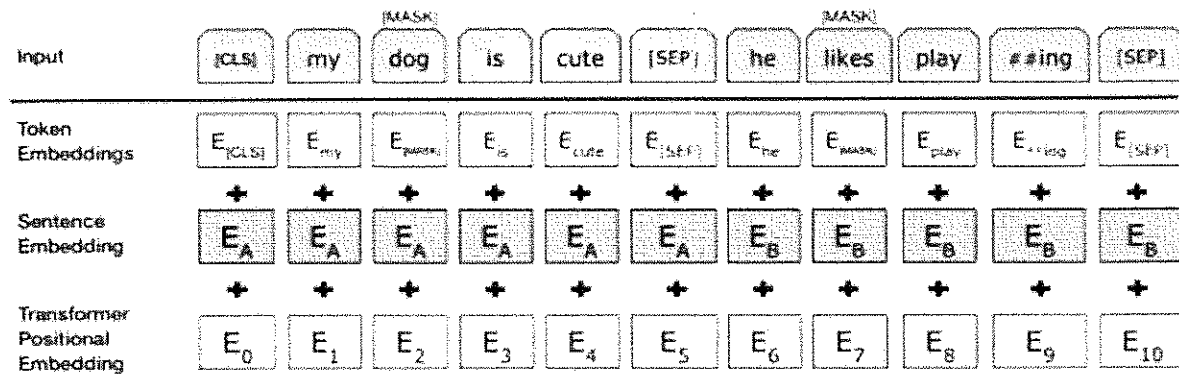


Figure 18:Source: BERT (Devlin, Chang, Lee, & Toutanova. 2018),with modifications

## 4. Experimental Setup

The main theme of our proposed solution is to assist psychologists in their manual time-consuming tasks for screening depression and identifying subjects that have symptoms of clinical depression with different severity levels. Currently, when the patient visits the clinic, he/she fills a manually-designed depression detection questionnaire. And normally it has human error either by doing it without attention or the patient would like to hide something since the illness that is related to mental illness has huge stigma especially in our society. According to this, there is a huge need to build an automatic system that helps the specialists to detect if the patient has depression or not, and if there is depression what is its severity level.

Our system is built based on a Machine learning algorithm that learns from the social media posts of the users and predicts the result of the depression questionnaire. Normally, existing systems provide a probability of a user's depression, however our system will provide the actual severity category of a user's depression level based on material submitted by users utilizing natural language processing algorithms.

To achieve this objective and develop the proposed system's prototype, we used Google's colab with Python programming language to build up the model as an environment. We also used the

sklearn library to help in the construction of the natural language processing pipeline. For the performance check of the models, we have used the confusion matrix that shows the precision recall, accuracy and F1 scores.

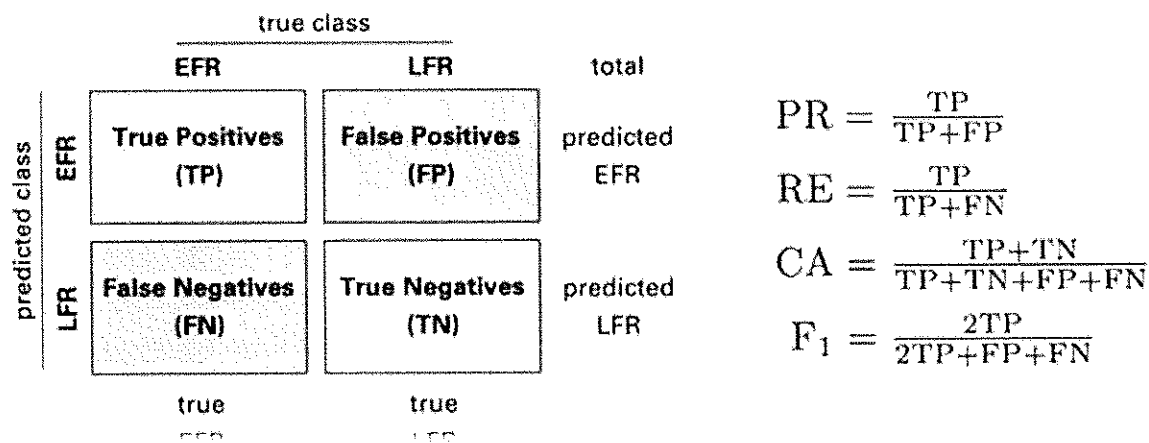


Figure 19: confusion matrix

Using the training data that have been provided by eRisk CLEF 2020, each occurrence of this dataset was characterized by the highlights specified in section 4.1 and linked to one user. For each user, the highlights were extracted from the sequentially concatenated works of all their chunks. For each model, we give the same data splitting between training and test to check the performance and we apply the classifiers that we have mentioned in the machine learning section of this thesis and get the values of the performance to compare the result to give the best result to system users.

## 4.1 Machine Learning Parameters Optimization

During the coding process of the project, we used different algorithms of Machine learning to predict the label of the social media posts. As we previously mentioned, we have used logistic regression, Naïve Bayes, SVM, RF, BERT, and XGBoost for the English language dataset. On the other hand, for the Arabic language dataset, we used part of these algorithms like the logistic regression, naïve Bayes, RF, and the SVM. To improve the testing accuracy of the system we used different optimizer techniques to see the effects of changing the parameters over the final accuracy of the algorithm, next section will summarize the results of the optimization parameters. Usually, the default parameters of the solver work well over most of the scenarios. But still there is a need to work over the solver of the optimizer in some cases like the huge dataset and the dataset that is built to perform prediction over huge datasets of text. The default solver of the classifiers is the ibfgs which is called Limited-memory BFGS. This solver just computes a Gaussian approximation based on the gradient, which makes it more computationally expensive. Furthermore, as compared to conventional bfgs, its memory usage is limited, causing it to reject older gradients and accumulate only new gradients as allowed by the memory constraint. newton-cg: Solver for explicitly computing Hessian, which can be computationally expensive in high dimensions. Stochastic Average Gradient Descent is abbreviated as sag. With enormous datasets, a more efficient solver is needed. Saga is a Sag variant that can be combined with l1 Regularization. It's a fast solver that's frequently used when dealing with very large datasets (Jaakkola & Jordan, 1997).

## 4.2 Discussion and Analysis of Findings

We used six classifiers over the training dataset. The training set has different variables that have an effect on the performance of the machine learning models. First, we started with the logistic regression algorithms which shows accuracy of 0.58 with the default values of the function but once we change the parameters the accuracy is fluctuating based on the table below:

*Table 2: Logistic regression Accuracy*

Number of tests	Ngram	LF solver	Accuracy
1.	(1,2)	Saga	0.558
2.	(1,3)	Saga	0.542
3.	(1,4)	saga	0.531
4.	(1,4)	lbfgs	0.532
5.	(1,2)	lbfgs	0.546
6.	(1,2)	newton-cg	0.558
7.	(1,2)	sag	0.558

For the Naïve bayes, we have tried the function over the default parameters values and after applying some parameters to check the performance change, Table 2 presents the results of the performance.

*Table 3: Naïve Bayes with n-gram change*

Number of tests	Ngram	Naïve -Accuracy
1.	default	0.503
2.	(1,2)	0.479
3.	(1,3)	0.478
4.	(1,4)	0.474



For the XGB boost, we have tried the default values and also the performance reflect after changing the ngrams values and the effect of changing the default parameters of tf-idf, the result is listed in the Table 4 below:

*Table 4: XGB boost Accuracy*

Number of tests	Ngram	boost solver	Accuracy
1.	Default	Default	0.471
2.	(1,2)	Saga	0.474
3.	(1,3)	saga	0.471
4.	(1,2)	lbfgs	0.474
5.	(1,3)	lbfgs	0.471
6.	(1,2)	newton-cg	0.474
7.	(1,2)	sag	0.474

While testing the model using a random forest classifier the default parameters of the model function have produced close results compared with the result that we achieved after giving specific n-grams parameters to function, Table 5 shows the accuracy performance over the random forest classifier.

*Table 5: Random Forest classifier*

Number of tests	Ngram	RF -Accuracy
1.	default	0.49
2.	(1,2)	0.489
3.	(1,3)	0.489

As per the work done before, and the results shown, we have noticed that models are showing low accuracy and we started studying how can we improve the model's accuracy, so we designed a preprocessing function that has the cleaning text, removing stop words, calculate the tf-idf based on unigrams. The results of ~~the~~ ~~are~~ ~~shown~~ are shown in the tables below.

The pre-processing steps that ~~were~~ ~~used~~ ~~are~~ ~~shown~~ are shown

Text cleaning: which is the stage where we remove the special characters, numbers, and punctuations.

Pre-processing the corpus: which applies the same rule in the dataset and the features.

Encoding the labels: in our case we do have four main classes for the depression severity which are the mild, minimal, severe and moderate.

Compute the tf-idf: which will be removing the English stop words and use the unigram before calculating the tf-idf.

For the Naïve base model, the best accuracy was 0.503 while after applying the text cleaning and preprocessing stages we got the confusion matrix below.

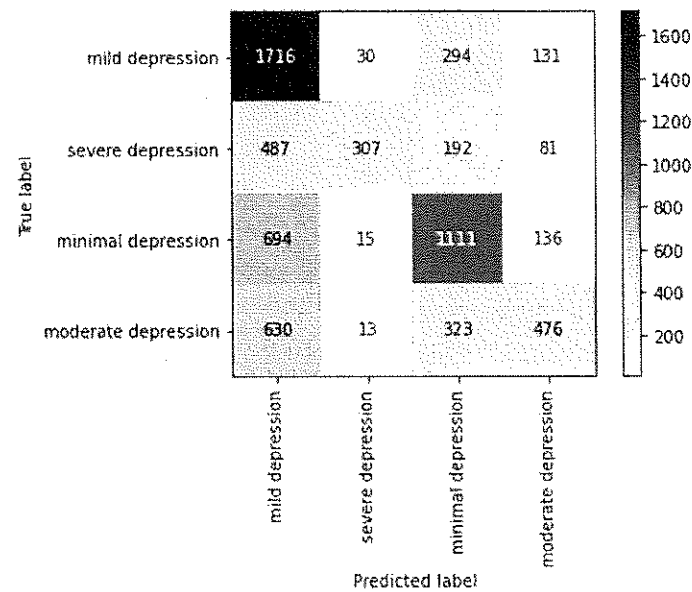
Table 6: Confusion matrix of Naïve model

Number of text = 26542   Number of features = 16105				
Start training...done!				
Start testing...done!				
Total time: 0.05s				
accuracy: 0.5440024110910187				
	precision	recall	f1-score	support
mild depression	0.49	0.79	0.60	2171
severe depression	0.84	0.29	0.43	1067
minimal depression	0.58	0.57	0.57	1956
moderate depression	0.58	0.33	0.42	1442
accuracy			0.54	6636
macro avg	0.62	0.49	0.51	6636
weighted avg	0.59	0.54	0.53	6636

After applying the pre-processing phase, we got an ~~improvement~~ over the Naïve Bayes model. It was 0.503 And it becomes 0.544 with ~~improvement~~ ~~of each depression severity~~ label. Table 7 shows the confusion matrix of ~~the~~ ~~with~~ the details of each depression

severity. And Table 8 presents the predicted values of each depression severity using the Naïve Bayes model.

Table 7: Predicted outcomes of depression severity - Naïve bayes



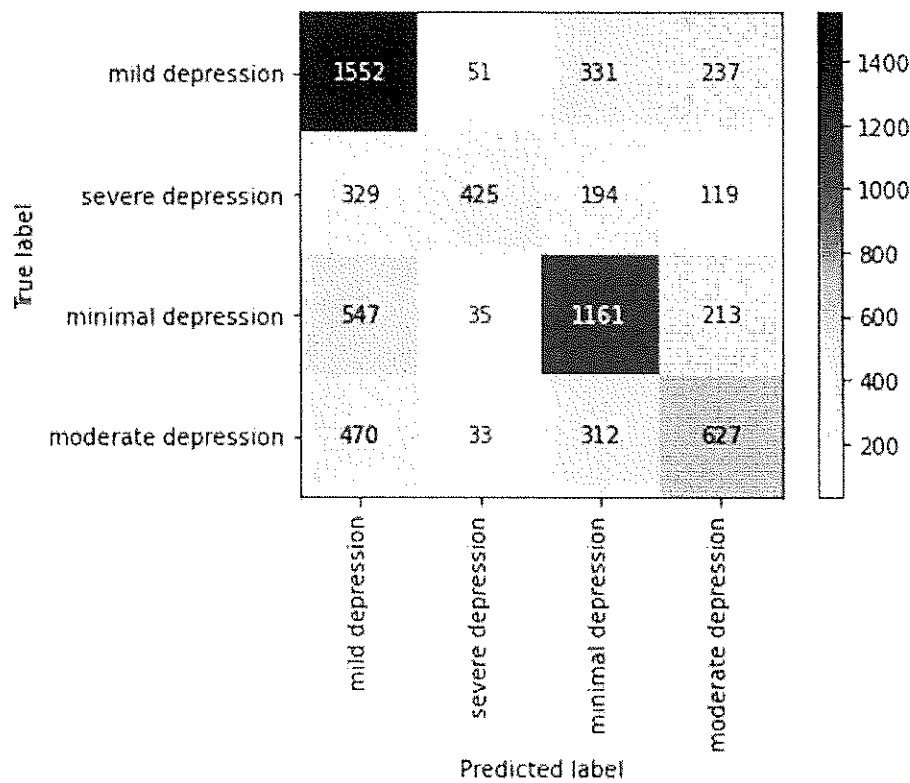
Next, for the Logistic regression, after applying the pre-processing and cleaning stages over the row dataset we have noticed improvement over the accuracy of the model and improvement over the F1 score of each of the depression severity categories. Table 8 gives the confusion matrix of the model.

Table 8: Logistic Regression Confusion Matrix

Number of documents = 26542   Number of features = 16105				
Start training...done!				
Start testing...done!				
Total time: 11.56s				
accuracy: 0.5673598553345389				
	precision	recall	f1-score	support
mild depression	0.54	0.71	0.61	2171
severe depression	0.78	0.40	0.53	1067
minimal depression	0.58	0.59	0.59	1956
moderate depression	0.52	0.43	0.48	1442
accuracy			0.57	6636
macro avg	0.61	0.54	0.55	6636
weighted avg	0.59	0.57	0.56	6636

The accuracy of the Logistic Regression model was 0.558 while after applying the pre-processing phases the accuracy improved to become 0.57 and the F1 score improved over the depression severity. The predicted outcomes shown in Table 9 per reflect the various severity levels predicted by the utilized model.

Table 9: Predicted outcomes of LR



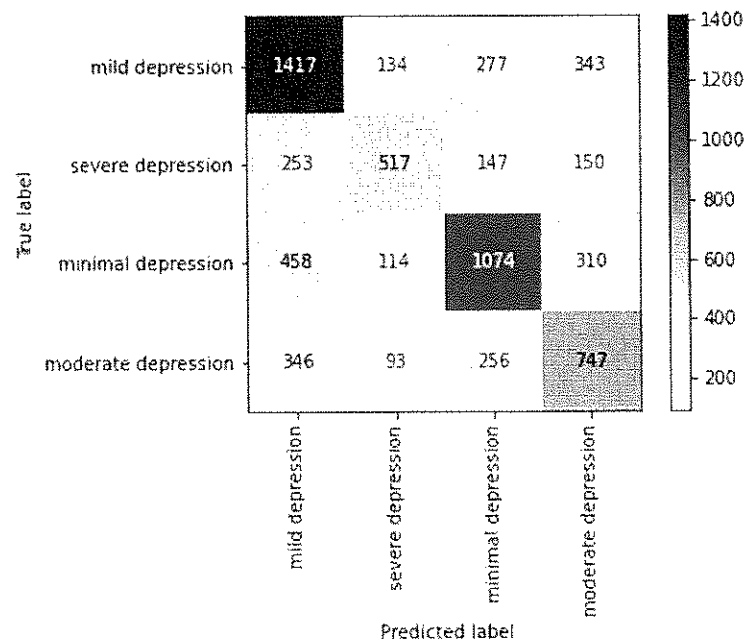
We applied the support vector machine over the cleaned and preprocessed dataset, and the accuracy has shown better results compared with the other classifiers that have been built before. Table 10 shows the confusion matrix of the SVM.

Table 10: confusion matrix of SVM

Number of documents = 26542   Number of features = 16105				
Start training...done!				
Start testing...done!				
Total time: 0.25s				
accuracy: 0.5658529234478602				
	precision	recall	f1-score	support
mild depression	0.57	0.65	0.61	2171
severe depression	0.60	0.48	0.54	1067
minimal depression	0.61	0.55	0.58	1956
moderate depression	0.48	0.52	0.50	1442
accuracy			0.57	6636
macro avg	0.57	0.55	0.56	6636
weighted avg	0.57	0.57	0.57	6636

The accuracy of the SVM is 0.61. While the prediction outcomes over the depression severity category shown in the Table below:

Table 11: SVM Predict outcomes



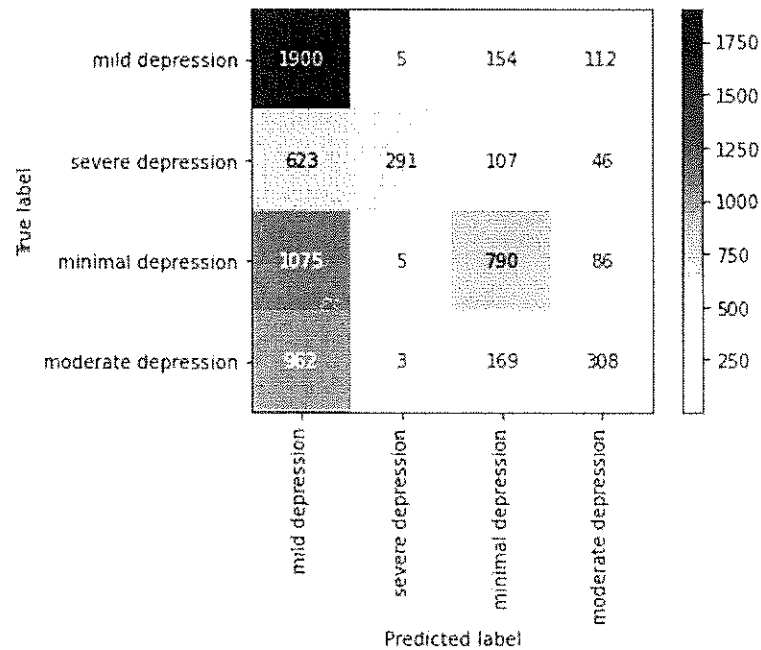
Random Forest classifier on the other hand is not showing improvement over the accuracy even after applying the pre-processing and cleaning phases. The model was showing 0.49 and now after applying it showing 0.50, the Table 11 below which depicts the confusion matrix.

Table 12: confusion matrix of Random Forest

Number of documents = 26542   Number of features = 16105				
Start training...done!				
Start testing...done!				
Total time: 9.62s				
accuracy: 0.4956298975286317				
	precision	recall	f1-score	support
mild depression	0.42	0.88	0.56	2171
severe depression	0.96	0.27	0.42	1067
minimal depression	0.65	0.40	0.50	1956
moderate depression	0.56	0.21	0.31	1442
accuracy			0.50	6636
macro avg	0.64	0.44	0.45	6636
weighted avg	0.60	0.50	0.47	6636

And the predicted outcomes based on the model are illustrated in Table 13.

Table 13: Random Forest predicted outcomes



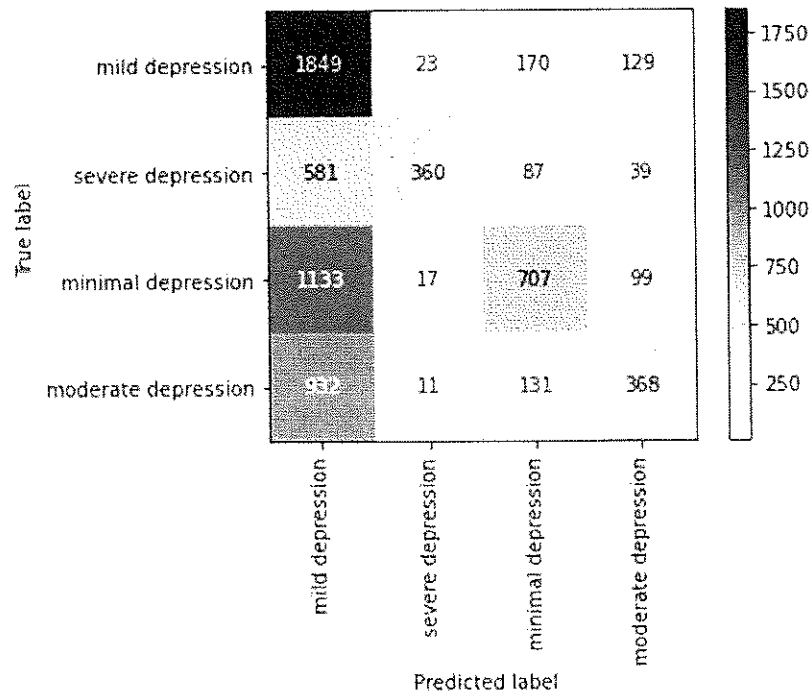
As we mentioned before, the boosting algorithm has been used and it was showing accuracy of 0.474 and after applying the cleaning phase it has been shown 0.49 as per the confusion matrix below.

Table 14: Confusion matrix of Gradient boosting

Start training...done!				
Start testing...done!				
accuracy: 0.4948764315852923				
	precision	recall	f1-score	support
mild depression	0.41	0.85	0.55	2171
severe depression	0.88	0.34	0.49	1067
minimal depression	0.65	0.36	0.46	1956
moderate depression	0.58	0.26	0.35	1442
accuracy			0.49	6636
macro avg	0.63	0.45	0.46	6636
weighted avg	0.59	0.49	0.47	6636

The predicted outcomes of the gradient boosting shown in the Table 15 below.

Table 15: Gradient boosting predicted outcomes



For the BERT Analysis, we have used the pre-trained model by BERT to train the data with 5 epochs and the accuracy of model starts from 57% and ends with 87% on the last training stage.



This Model is taking time to train the model over the dataset especially that the dataset that we have is huge and the training must go through 80% of the dataset. Also, the training is consuming processing power according to the huge computation that happen over the data content to fetch the useful information and explore the result.

```
tokenizer = BertTokenizer.from_pretrained('bert-base-cased')  
  
Downloading vocab.txt: 100% ██████████ 208k/208k [00:00<00:00, 621kB/s]  
Downloading tokenizer_config.json: 100% ██████████ 29.0/29.0 [00:00<00:00, 743B/s]  
Downloading config.json: 100% ██████████ 570/570 [00:00<00:00, 19.1kB/s]
```

BERT 1:Tokenizer

```

1 sentiment_model.compile(optimizer=optim, loss=loss_func, metrics=[acc])
hist = sentiment_model.fit(
    train_dataset,

    validation_data=val_dataset,
    epochs=5
)

Epoch 1/5
1658/1658 [=====] - 1467s 878ms/step - loss: 1.1110 - accuracy: 0.5125 - val_loss: 0.9707 - val_accuracy: 0.5798
Epoch 2/5
1658/1658 [=====] - 1453s 876ms/step - loss: 0.9225 - accuracy: 0.6164 - val_loss: 0.8255 - val_accuracy: 0.6634
Epoch 3/5
1658/1658 [=====] - 1453s 876ms/step - loss: 0.7683 - accuracy: 0.6968 - val_loss: 0.6787 - val_accuracy: 0.7393
Epoch 4/5
1658/1658 [=====] - 1453s 876ms/step - loss: 0.6044 - accuracy: 0.7661 - val_loss: 0.5223 - val_accuracy: 0.8117
Epoch 5/5
1658/1658 [=====] - 1453s 876ms/step - loss: 0.4570 - accuracy: 0.8299 - val_loss: 0.3914 - val_accuracy: 0.8670

```

BERT 2: BERT Accuracy

For the Arabic dataset, we have built several techniques using various Machine learning models to predict the depression status out of the social media posts. We started the work over the Arabic dataset by implementing the tokenization over the text using several techniques.

- 1) Tokenization: in python we use the split function which is a built in function in python.

```
df2['split']=df1['Text'].apply(lambda x: str.split(x))
df2
```

	Text	Label	split
1	...البطزين يحيى من السعوديه بأقل الاسعار و الحكومة	NEG	[...البطزين يحيى من السعوديه بأقل الاسعار و الحكومة]
2	الله يوفقكم	NEG	[الله يوفقكم]
3	...لأسف البعض يعتقد ان المفاعل النووي سيحل مشاكل	NEG	[...لأسف البعض يعتقد ان المفاعل النووي سيحل مشاكل]
4	...المفاعل النووي يحتاج لشريك استراتيجي يمتلكه ا	NEG	[...المفاعل النووي يحتاج لشريك استراتيجي يمتلكه ا]
6	...المفاعل النووي يحتاج لشريك استراتيجي يمتلكه ا	NEG	[...المفاعل النووي يحتاج لشريك استراتيجي يمتلكه ا]
...	...	...	...
396	أكد لوفى كده الدنيا	POS	[أكد لوفى كده الدنيا]
397	الله يوفقك	POS	[الله يوفقك]
398	و الله ثق نفسك ده يقول كلام زى الفل	POS	[و الله ثق نفسك ده يقول كلام زى الفل]
399	كل الاحترام	POS	[كل الاحترام]
400	نعم	POS	[نعم]

Figure 20: Python split function

## 2) Wordnet Tokenizer:

```
df2["wordnet_tokenizer"]=df2.apply(lambda row: nltk.word_tokenize(row["Text"]), axis=1)
df2
```

	Text	Label	split	wordnet_tokenizer
1	...البطزين يحيى من السعوديه بأقل الاسعار و الحكومة	NEG	[...البطزين يحيى من السعوديه بأقل الاسعار و الحكومة]	[...البطزين يحيى من السعوديه بأقل الاسعار و الحكومة]
2	الله يوفقكم	NEG	[الله يوفقكم]	[الله يوفقكم]
3	...لأسف البعض يعتقد ان المفاعل النووي سيحل مشاكل	NEG	[...لأسف البعض يعتقد ان المفاعل النووي سيحل مشاكل]	[...لأسف البعض يعتقد ان المفاعل النووي سيحل مشاكل]
4	...المفاعل النووي يحتاج لشريك استراتيجي يمتلكه ا	NEG	[...المفاعل النووي يحتاج لشريك استراتيجي يمتلكه ا]	[...المفاعل النووي يحتاج لشريك استراتيجي يمتلكه ا]
6	...المفاعل النووي يحتاج لشريك استراتيجي يمتلكه ا	NEG	[...المفاعل النووي يحتاج لشريك استراتيجي يمتلكه ا]	[...المفاعل النووي يحتاج لشريك استراتيجي يمتلكه ا]
...	...	...	...	...
396	أكد لوفى كده الدنيا	POS	[أكد لوفى كده الدنيا]	[أكد لوفى كده الدنيا]
397	الله يوفقك	POS	[الله يوفقك]	[الله يوفقك]
398	و الله ثق نفسك ده يقول كلام زى الفل	POS	[و الله ثق نفسك ده يقول كلام زى الفل]	[و الله ثق نفسك ده يقول كلام زى الفل]
399	كل الاحترام	POS	[كل الاحترام]	[كل الاحترام]
400	نعم	POS	[نعم]	[نعم]

Figure 21: wordnet Tokenizer

Also in the second phase we performed the stemming over the text to see the Arabic words that have been used by the users. Figure 22 illustrates the code and the output when using the Lancaster stemmer:

```
def lanc(words):
    lanca = LancasterStemmer()
    return [lanca.stem(word) for word in words.split()]
df2['Lanc_stemmer']=df2['Text'].apply(lambda x: lanc(x))
df2
```

	Text	Label	split	wordnet_tokenizer	Lanc_stemmer
1	...الذين يحيى من السعوديه بأقل الاسعار و الحكومه	NEG	...الذين يحيى من السعوديه بأقل الاسعار و ...	...الذين يحيى من السعوديه بأقل الاسعار و ...	...الذين يحيى من السعوديه بأقل الاسعار و ...
2	الله يرحمك	NEG	[الله يرحمك]	[الله يرحمك]	[الله يرحمك]
3	...أشرف المهن بفتح الـ المفاعل النووي سيجب مشاكل	NEG	...أشرف المهن بفتح الـ المفاعل النووي سيجب ...	...أشرف المهن بفتح الـ المفاعل النووي سيجب ...	...أشرف المهن بفتح الـ المفاعل النووي سيجب ...
4	...المفاعل النووي سيجب اشراك استراتيجي يمكنه	NEG	...المفاعل النووي سيجب اشراك استراتيجي يمكنه ...	...المفاعل النووي سيجب اشراك استراتيجي يمكنه ...	...المفاعل النووي سيجب اشراك استراتيجي يمكنه ...
6	...المفاعل النووي سيجب اشراك استراتيجي يمكنه	NEG	...المفاعل النووي سيجب اشراك استراتيجي يمكنه ...	...المفاعل النووي سيجب اشراك استراتيجي يمكنه ...	...المفاعل النووي سيجب اشراك استراتيجي يمكنه ...
...	...	...	...	...	...
396	أعيد فر في كده الشيا	POS	[أعيد فر في كده الشيا]	[أعيد فر في كده الشيا]	[أعيد فر في كده الشيا]
397	الله يرحمك	POS	[الله يرحمك]	[الله يرحمك]	[الله يرحمك]
398	و الله لك نفسك ده يقول كده في الف	POS	و الله لك نفسك ده يقول كده في الف	و الله لك نفسك ده يقول كده في الف	و الله لك نفسك ده يقول كده في الف
399	كنا الاحرام	POS	[كنا الاحرام]	[كنا الاحرام]	[كنا الاحرام]
400	نعم	POS	[نعم]	[نعم]	[نعم]

Figure 22 Lancaster lemmatizer

We have also used the ISRI stemmer to check what will be the best function to be used in our work as depicted in the below figure.

```
def stemming(words):
    st = ISRIStemmer()
    return [st.stem(word) for word in words.split()]
df2['Isri_stemmer']=df2['Text'].apply(lambda x: stemming(x))
df2
```

	Text	Label	split	wordnet_tokenizer	Lanc_stemmer	Isri_stemmer
1	...الذين يحيى من السعوديه بأقل الاسعار و الحكومه	NEG	...الذين يحيى من السعوديه بأقل الاسعار و ...	...الذين يحيى من السعوديه بأقل الاسعار و ...	...الذين يحيى من السعوديه بأقل الاسعار و ...	...الذين يحيى من السعوديه بأقل الاسعار و ...
2	الله يرحمك	NEG	[الله يرحمك]	[الله يرحمك]	[الله يرحمك]	[الله يرحمك]
3	...أشرف المهن بفتح الـ المفاعل النووي سيجب مشاكل	NEG	...أشرف المهن بفتح الـ المفاعل النووي سيجب ...	...أشرف المهن بفتح الـ المفاعل النووي سيجب ...	...أشرف المهن بفتح الـ المفاعل النووي سيجب ...	...أشرف المهن بفتح الـ المفاعل النووي سيجب ...
4	...المفاعل النووي سيجب اشراك استراتيجي يمكنه	NEG	...المفاعل النووي سيجب اشراك استراتيجي يمكنه ...	...المفاعل النووي سيجب اشراك استراتيجي يمكنه ...	...المفاعل النووي سيجب اشراك استراتيجي يمكنه ...	...المفاعل النووي سيجب اشراك استراتيجي يمكنه ...
6	...المفاعل النووي سيجب اشراك استراتيجي يمكنه	NEG	...المفاعل النووي سيجب اشراك استراتيجي يمكنه ...	...المفاعل النووي سيجب اشراك استراتيجي يمكنه ...	...المفاعل النووي سيجب اشراك استراتيجي يمكنه ...	...المفاعل النووي سيجب اشراك استراتيجي يمكنه ...
...	...	...	...	...	...	...
396	أعيد فر في كده الشيا	POS	[أعيد فر في كده الشيا]	[أعيد فر في كده الشيا]	[أعيد فر في كده الشيا]	[أعيد فر في كده الشيا]
397	الله يرحمك	POS	[الله يرحمك]	[الله يرحمك]	[الله يرحمك]	[الله يرحمك]
398	و الله لك نفسك ده يقول كده في الف	POS	و الله لك نفسك ده يقول كده في الف	و الله لك نفسك ده يقول كده في الف	و الله لك نفسك ده يقول كده في الف	و الله لك نفسك ده يقول كده في الف
399	كنا الاحرام	POS	[كنا الاحرام]	[كنا الاحرام]	[كنا الاحرام]	[كنا الاحرام]
400	نعم	POS	[نعم]	[نعم]	[نعم]	[نعم]

Figure 23: ISRI stemmer

Then we have removed the stopwords of the Arabic language to make sure that all the words that will enter as input to the model have a contribution to the meaning of the text. Figure 24 shows the output of the stopwords removal function of using the NLTK library:

	Text	Label	split	wordnet_tokenizer	Lanc_stemmer	Isri_stemmer	no_stopwords
1	الذين يحيى من السموم يقاتل ...الاسماء والحكومة	NEG	[الذين يحيى من السموم يقاتل ...الاسماء والحكومة]	[الذين يحيى من السموم يقاتل ...الاسماء والحكومة]	[الذين يحيى من السموم يقاتل ...الاسماء والحكومة]	[الذين يحيى من السموم يقاتل ...الاسماء والحكومة]	[الذين يحيى من السموم يقاتل ...الاسماء والحكومة]
2	انه يوضح	NEG	[انه يوضح]	[انه يوضح]	[انه يوضح]	[انه يوضح]	[انه يوضح]
3	لشخص البعض يعتقد ان المقاتل ...الوطني يوضح	NEG	[لشخص البعض يعتقد ان المقاتل ...الوطني يوضح]	[لشخص البعض يعتقد ان المقاتل ...الوطني يوضح]	[لشخص البعض يعتقد ان المقاتل ...الوطني يوضح]	[لشخص البعض يعتقد ان المقاتل ...الوطني يوضح]	[لشخص البعض يعتقد ان المقاتل ...الوطني يوضح]
4	المقاتل القومي يحتاج لشريك ...الوطني يوضح	NEG	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]
6	المقاتل القومي يحتاج لشريك ...الوطني يوضح	NEG	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]
...	...	...	...	...	...	...	...
396	كثير من كذا	POS	[كثير من كذا]	[كثير من كذا]	[كثير من كذا]	[كثير من كذا]	[كثير من كذا]
397	انه يوضح	POS	[انه يوضح]	[انه يوضح]	[انه يوضح]	[انه يوضح]	[انه يوضح]
398	و انه تفكك كذا ...الوطني يوضح	POS	[و انه تفكك كذا ...الوطني يوضح]	[و انه تفكك كذا ...الوطني يوضح]	[و انه تفكك كذا ...الوطني يوضح]	[و انه تفكك كذا ...الوطني يوضح]	[و انه تفكك كذا ...الوطني يوضح]
399	كذا	POS	[كذا]	[كذا]	[كذا]	[كذا]	[كذا]
400	...	POS	[...]	[...]	[...]	[...]	[...]

Figure 24: Arabic nltk stopwords removal

Then we lemmatized the text using WordNet lemmatizer and the result are shown in the Figure 25 below:

```
df2['wordnet_lemmatizer']=df2['Text'].apply(lambda x: lemma(x))
df2
```

	Text	Label	split	wordnet_tokenizer	Lanc_stemmer	Isri_stemmer	no_stopwords	wordnet_lemmatizer
1	الذين يحيى من السموم يقاتل ...الاسماء والحكومة	NEG	[الذين يحيى من السموم يقاتل ...الاسماء والحكومة]	[الذين يحيى من السموم يقاتل ...الاسماء والحكومة]	[الذين يحيى من السموم يقاتل ...الاسماء والحكومة]	[الذين يحيى من السموم يقاتل ...الاسماء والحكومة]	[الذين يحيى من السموم يقاتل ...الاسماء والحكومة]	[الذين يحيى من السموم يقاتل ...الاسماء والحكومة]
2	انه يوضح	NEG	[انه يوضح]	[انه يوضح]	[انه يوضح]	[انه يوضح]	[انه يوضح]	[انه يوضح]
3	لشخص البعض يعتقد ان المقاتل ...الوطني يوضح	NEG	[لشخص البعض يعتقد ان المقاتل ...الوطني يوضح]	[لشخص البعض يعتقد ان المقاتل ...الوطني يوضح]	[لشخص البعض يعتقد ان المقاتل ...الوطني يوضح]	[لشخص البعض يعتقد ان المقاتل ...الوطني يوضح]	[لشخص البعض يعتقد ان المقاتل ...الوطني يوضح]	[لشخص البعض يعتقد ان المقاتل ...الوطني يوضح]
4	المقاتل القومي يحتاج لشريك ...الوطني يوضح	NEG	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]
6	المقاتل القومي يحتاج لشريك ...الوطني يوضح	NEG	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]	[المقاتل القومي يحتاج لشريك ...الوطني يوضح]
...	...	...	...	...	...	...	...	...
396	كثير من كذا	POS	[كثير من كذا]	[كثير من كذا]	[كثير من كذا]	[كثير من كذا]	[كثير من كذا]	[كثير من كذا]
397	انه يوضح	POS	[انه يوضح]	[انه يوضح]	[انه يوضح]	[انه يوضح]	[انه يوضح]	[انه يوضح]
398	و انه تفكك كذا ...الوطني يوضح	POS	[و انه تفكك كذا ...الوطني يوضح]	[و انه تفكك كذا ...الوطني يوضح]	[و انه تفكك كذا ...الوطني يوضح]	[و انه تفكك كذا ...الوطني يوضح]	[و انه تفكك كذا ...الوطني يوضح]	[و انه تفكك كذا ...الوطني يوضح]
399	كذا	POS	[كذا]	[كذا]	[كذا]	[كذا]	[كذا]	[كذا]
400	...	POS	[...]	[...]	[...]	[...]	[...]	[...]

Figure 25: WordNet Lemmatizer

We have tried the qlassdi ~~XXXXXXXXXXXXXXXXXXXX~~ in Figure 26 below:

```
def lemma(words):
    lm = qalsadi.Lemmatizer.Lemmatizer()
    return [lm.lemmatize(word) for word in words.split()]

df2['qalsadi_lemmatizer']=df2['Text'].apply(lambda x: lemma(x))
df2
```

	Text	Label	split	wordnet_tokenizer	Lanc_stemmer	iso_stemmer	no_stopwords	wordnet_lemmatizer	qalsadi_lemmatizer
1	الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...	NEG	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]
2	الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...	NEG	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]
3	الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...	NEG	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]
4	الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...	NEG	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]
6	الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...	NEG	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]
396	الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...	POS	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]
397	الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...	POS	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]	[الذين يبيعون السيارات يبيعون السيارات و السيارات و السيارات و...]

Figure 26: Qalsadi Lemmatizer

After preparing the text, we employed the Machine learning algorithms starting the Logistic regression model as demonstrated in the Figure 27 below:

```
#logistic Regression
pipe = make_pipeline(TfidfVectorizer(),
                      LogisticRegression())

# make param grid
param_grid = {'logisticregression__C': [0.01, 0.1, 1, 10, 100]}

# create and fit the model
model = GridSearchCV(pipe, param_grid, cv=5)
model.fit(X_train,Y_train)

# make prediction and print accuracy
prediction = model.predict(X_test)
print(f"Accuracy score is {accuracy_score(Y_test, prediction):.2f}")
print(classification_report(Y_test, prediction))

Accuracy score is 0.81
precision    recall  f1-score   support

0           0.79      0.82      0.81        38
1           0.83      0.81      0.82        42

accuracy          0.81
macro avg         0.81      0.81      0.81
weighted avg      0.81      0.81      0.81
```

Figure 27: Logistic Regression

When utilizing the Random Forest, an accuracy of 0.81 has been achieved when using the cleaned text after applying all the filters and removing the stopwords. Figure 28 shows a snippet of the developed code, as well as the results produced by the Random Forest model.

```
#Random Forest Classifier

pipe = make_pipeline(TfidfVectorizer(),
                    RandomForestClassifier())

param_grid = {'randomforestclassifier__n_estimators':[10, 100, 1000],
              'randomforestclassifier__max_features':['sqrt', 'log2']}

rf_model = GridSearchCV(pipe, param_grid, cv=5)
rf_model.fit(X_train,Y_train)

prediction = rf_model.predict(X_test)
print(f"Accuracy score is {accuracy_score(Y_test, prediction):.2f}")

Accuracy score is 0.70
```

Figure 28: Random Forest over Arabic Dataset

The Naïve Bayes model on the other hand shows 80% accuracy over the dataset, as depicted in the below figure.

```
#Naive Bayes Classifier

pipe = make_pipeline(TfidfVectorizer(),
                    MultinomialNB())
pipe.fit(X_train,Y_train)
prediction = pipe.predict(X_test)
print(f"Accuracy score is {accuracy_score(Y_test, prediction):.2f}")
print(classification_report(Y_test, prediction))

Accuracy score is 0.80
```

	precision	recall	f1-score	support
0	0.74	0.89	0.81	38
1	0.88	0.71	0.79	42
accuracy			0.80	80
macro avg	0.81	0.80	0.80	80
weighted avg	0.81	0.80	0.80	80

Figure 29: Naive Bayes

Next, using the support vector machine model, an accuracy rate of 81% was obtained. Figure 30 shows the performance matrix using this model.

```
#Support Vector Machine

pipe = make_pipeline(TfidfVectorizer(),
                    SVC())
param_grid = {'svc__kernel': ['rbf', 'linear', 'poly'],
              'svc__gamma': [0.1, 1, 10, 100],
              'svc__C': [0.1, 1, 10, 100]}

svc_model = GridSearchCV(pipe, param_grid, cv=5)
svc_model.fit(X_train, Y_train)

prediction = svc_model.predict(X_test)
print(f"Accuracy score is {accuracy_score(Y_test, prediction):.2f}")
print(classification_report(Y_test, prediction))

Accuracy score is 0.81
```

	precision	recall	f1-score	support
0	0.81	0.79	0.80	38
1	0.81	0.83	0.82	42
accuracy			0.81	80
macro avg	0.81	0.81	0.81	80
weighted avg	0.81	0.81	0.81	80

Figure 30: SVM of Arabic dataset

### 4.3 Visualizing the Significant Terms

In this section we would like to highlight the main significant terms that express the knowledge behind each of the depression severity. In this section, we presented the findings as tables to highlight the most frequently used terms by depressed individuals and to exhibit them as a word cloud.

As far as we know, the severity of depression in our dataset has been unequally distributed among the severities because the true data was gathered from social media postings written over the course of 2020. The percentage of each severity is shown in the pie chart below.

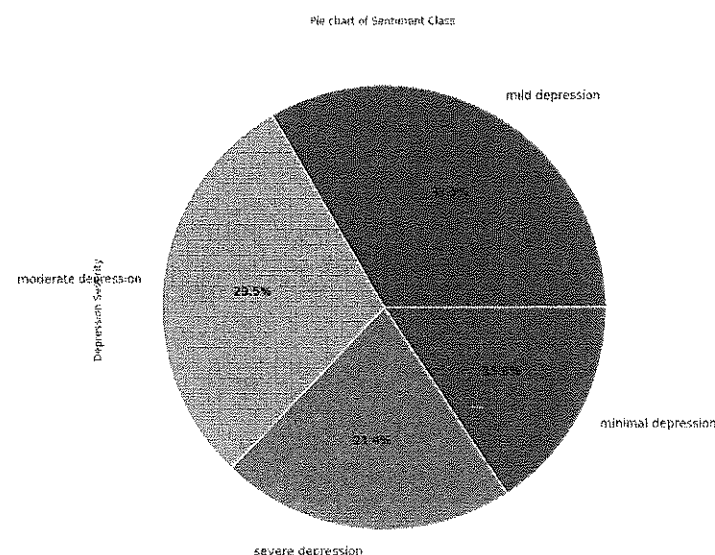


Figure 31: Depression Severity percentage

Also, we have built a variable called the `word_count` that counts the word occurrences in the dataset to know what are the most frequent words that have been used by the users when they write their posts over the internet. From this variable, we found that the top 20 words that have been used are listed on the table below.

	Common words	count
1	don't	5351
2	in	5308
3	one	5237
4	would	4518
5	people	4418
6	get	4232
7	think	3723
8	rule	3730
9	time	3508
10	comment	3569
11	thing	3305
12	know	3228
13	also	3036
14	make	2842
15	really	2807
16	see	2721
17	even	2673
18	want	2581
19	removed	2510

Figure 32: Top 20 terms that have been used by the users.

We also used the plotly library in python to show the words in a way that will be much easier for checking.

Tree of Most Common Words

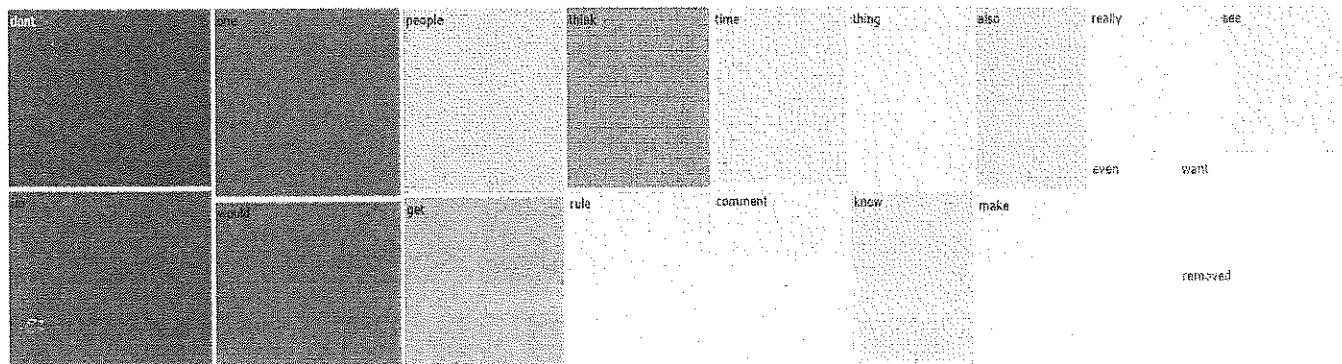


Figure 33: Plotly top 20 used terms

From the table and the plotly result that we have mentioned above, we noticed that term (don't) is listed. While it should be regarded as a stop word, however we should not remove it in this context. In other words, we don't remove it since the negation word can affect the ~~sentences~~ polarity of the sentences and because of this its effecting the depression severity as well.



Next, we depict the words and terms that have been used most frequently in each depression severity. We will start with the Mild depression. For the mild depression, we noticed that the users have used the term (Like) as the most frequent word which is 2343 times in the posts. The figure below shows the words and the counts of the mild depression posts.

Common Words in Mild depression

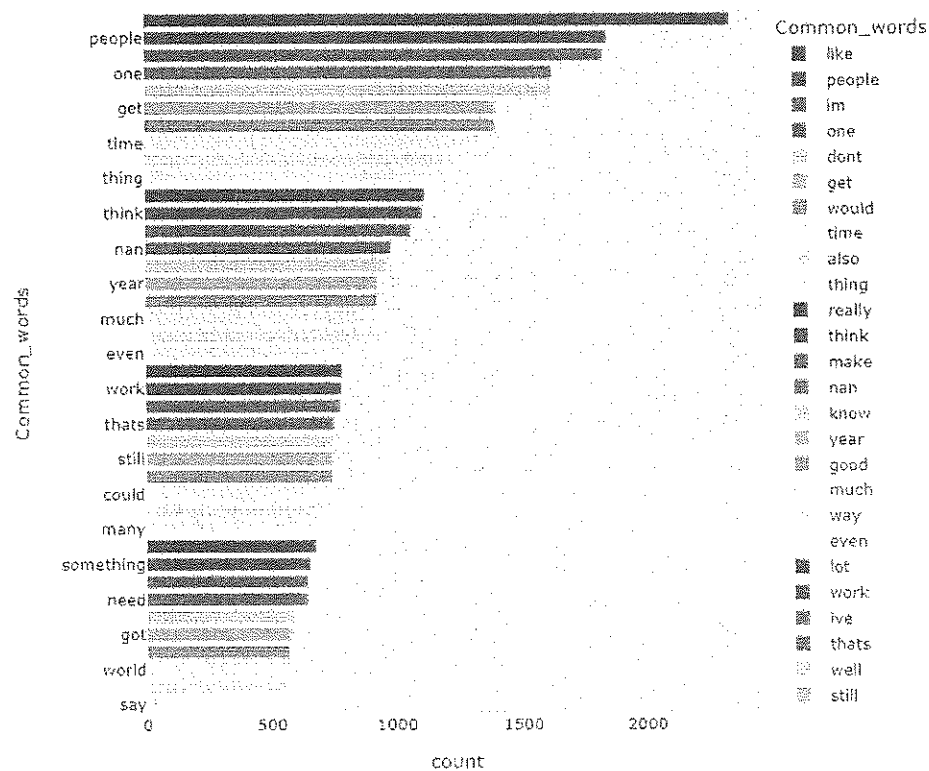


Figure 34:20 Mild Depression most frequent words

For the moderate depression posts, we have noticed that the term (Like) which has been mentioned 2084 times is the most used word in the moderate posts.

## Common Words in Moderate depression

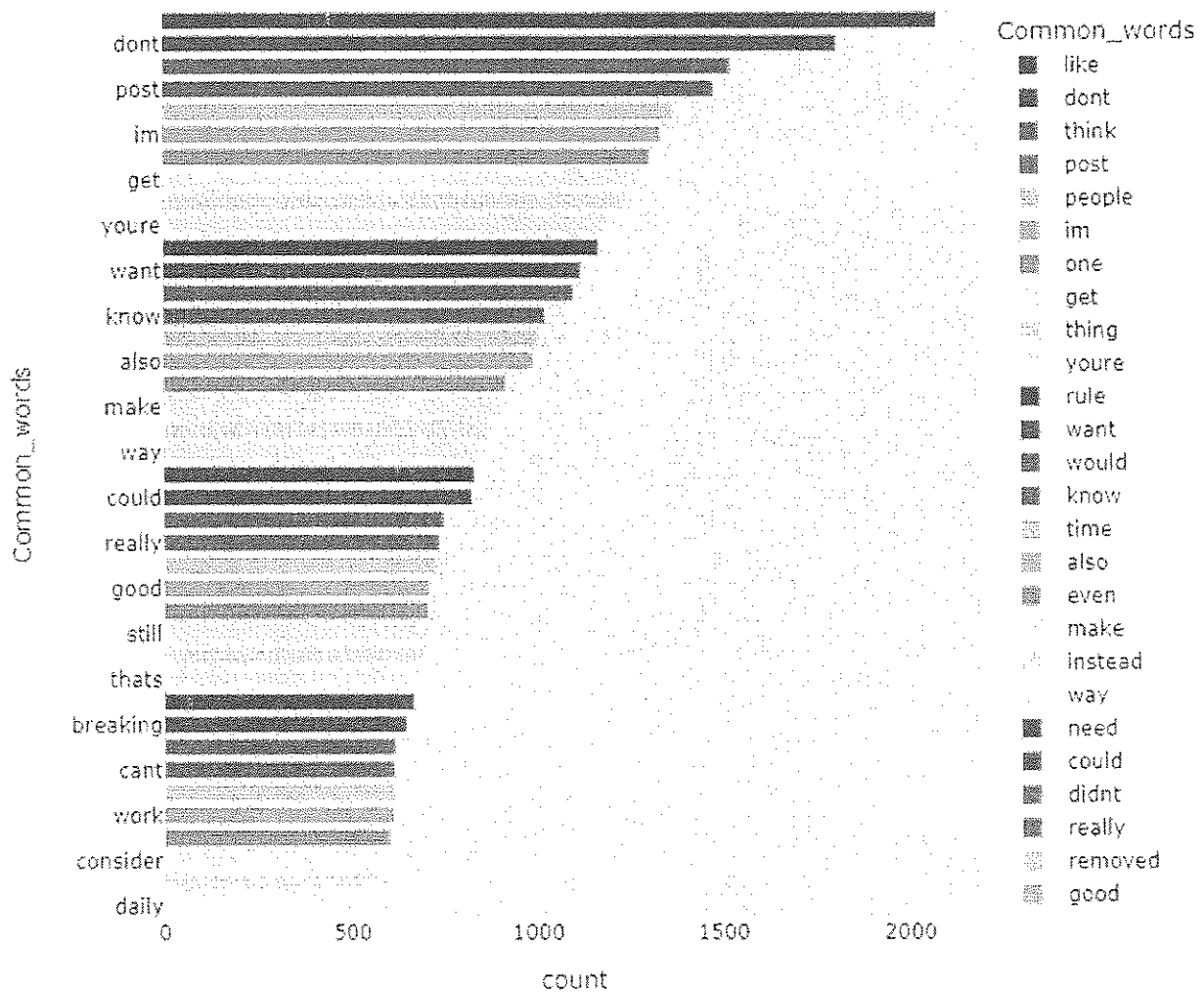


Figure 35: Moderate Depression most frequent words

For the minimal depression posts, the system is showing that most of the posts have the term (comment) and the number of counts for this term in the posts is 3052 times.

## Common Words in Minimal depression

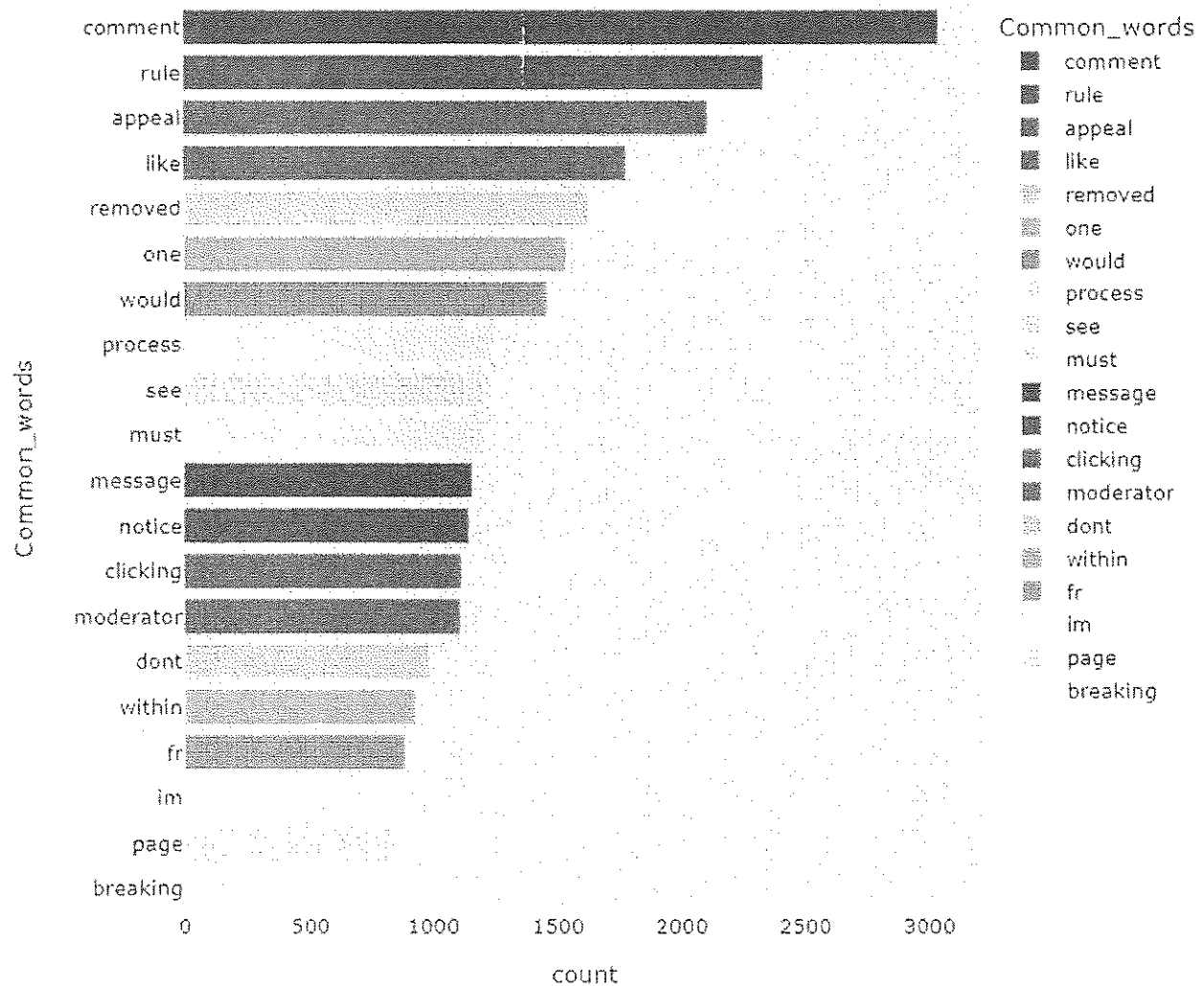


Figure 36: Minimal Depression most frequent words

The moderate depression has been also tested and we found that the most frequent term that has been used by the users on their social media posts is Hitler which is 1205 times.

## Common Words in severe depression

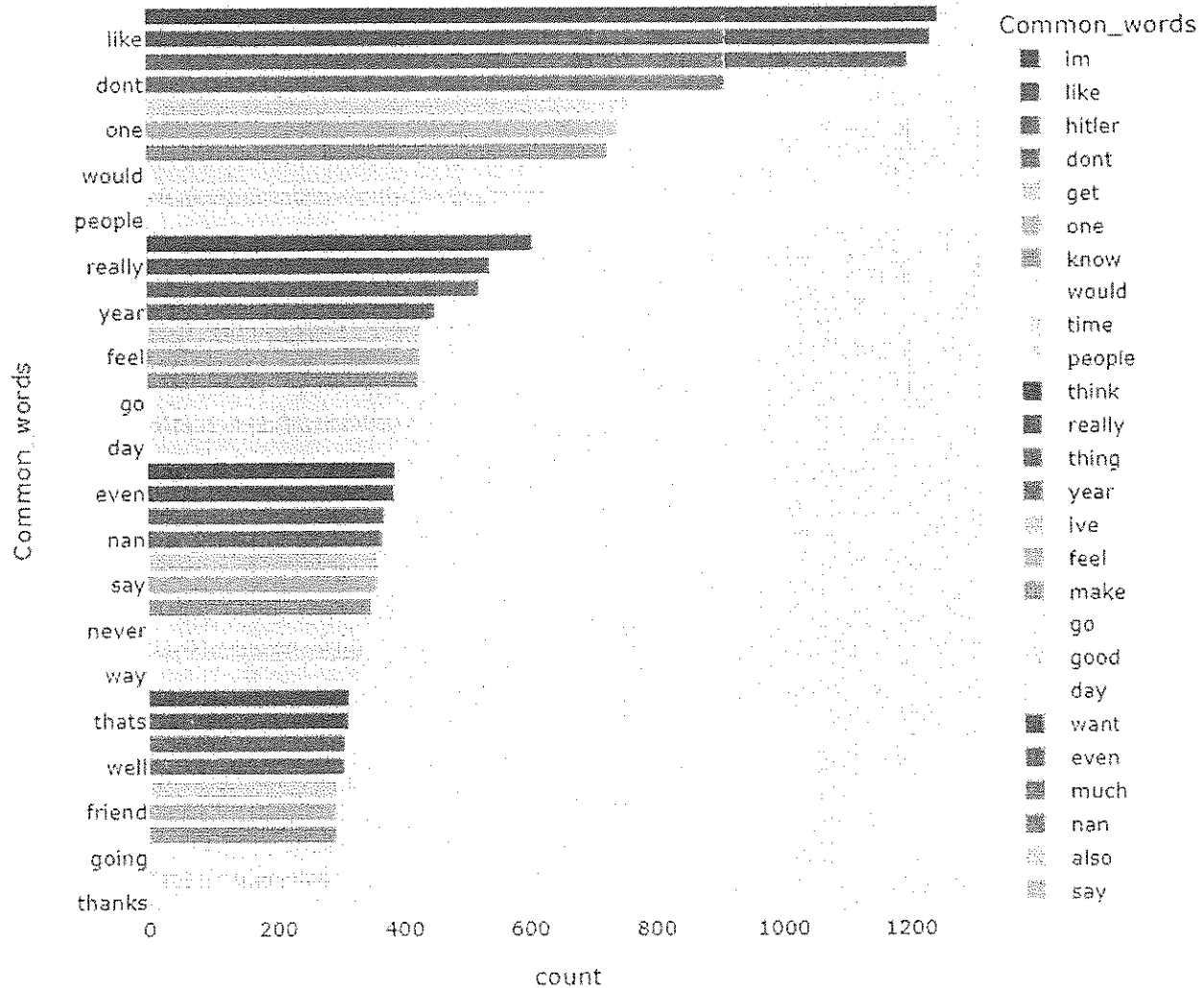


Figure 37: Severe depression most frequent words

From the analysis and the figures that have been shown before, we can notice that many terms are mutual between the posts that have been written by the users that have the same depression severity. So based on the posts that the user writes and the after studying the posts that have been posted by the user, we can have an idea about the depression severity that the user may have. It is important to mention here that we have collaborated with our colleagues in an ongoing research project that aims at the construction of a knowledge base that encodes the terms in the clinical

depression domain and constructs semantic networks that integrate multiple semantic resources in this context. We have provided our colleagues with significant terms that subjects identified with clinical depression use and express in their posts in an attempt to be incorporated as part of the proposed clinical depression knowledge base.

## **5. Conclusions and Recommendations**

The proposed machine learning models have examined and proved their ability to predict and recognize depression severity (using social media posts that users write to express their feeling and daily activities). We found that the incorporation of multiple techniques among the proposed NLP pipeline had a significant impact on the quality of the employed machine learning models. We experimentally investigated various scenarios using both English and Arabic datasets, and we found that despite the challenges inherent in each language, the analysis of social media posts can indeed be an alarming source that indicates whether a subject is suffering from depression and further recommend treatment procedures for identified subjects. This is particularly important since due to the stigma behind the mental illness in the society. As such, it was important to automate the process of screening depression since the current schools of screening use manual procedures that are based on depression questionnaires.

As we mentioned, the past work over the depression datasets were focusing on screening the depression with binary classifiers. In this context, the aim is to decide if the person has depression or not. However, in our system the model shows the depression severity of the ~~subject~~. Our proposed solution integrated BERT and produced promising accuracy results with ~~an~~ accuracy rate, which can be regarded as a reliable (still to be further improved) resource for ~~depression~~ screening.

It is also important to highlight the fact that utilization of the various datasets and their quality has had a huge impact on the quality of the produced results by each of the employed models. This was obviously noticed when dealing with the Arabic dataset that was skewed towards the depressed subjects. However, despite this fact, we argue that an improvement on the quality of the input training datasets can be performed to ensure producing a higher accuracy rate on the one hand, and more precise screening for subjects from both Arabic and English samples.

As a future extension to our currently proposed solution, we plan to acquire more datasets and involve real-world subjects from the Palestinian community to measure the effect of utilizing the proposed tool on our local community. We also plan to integrate more domain-specific clinical depression knowledge bases that can lead to improving the quality of feature engineering and consequently the machine learning process.

## References

- Abualigah, L., Alfar, H. E., Shehab, M., & Hussein, A. M. A. (2020). Sentiment analysis in healthcare: a brief review. *Recent Advances in NLP: The Case of Arabic Language*, 129-141.
- Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011). *Sentiment analysis of twitter data*. Paper presented at the Proceedings of the workshop on language in social media (LSM 2011).
- Aggarwal, C. C., & Wang, H. (2011). Text mining in social networks. In *Social network data analytics* (pp. 353-378): Springer.
- Al Asad, N., Pranto, M. A. M., Afreen, S., & Islam, M. M. (2019). *Depression detection by analyzing social media posts of user*. Paper presented at the 2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON).
- Al Shamsi, A. A., & Abdallah, S. (2021). Text mining techniques for sentiment analysis of Arabic dialects: literature review. *Adv. Sci. Technol. Eng. Syst. J*, 6, 1012-1023.
- Alajmi, A., Saad, E. M., & Darwish, R. (2012). Toward an ARABIC stop-words list generation. *International Journal of Computer Applications*, 46(8), 8-13.
- Alam, S., & Yao, N. (2019). The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. *Computational and Mathematical Organization Theory*, 22(3), 319-335.
- Almouzin, S., & Alkhatib, A. (2019). Detecting Arabic depressed users from Twitter data. *Proceedings of the 2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON)*, 163, 257-265.

- AlSagri, H. S., & Ykhlef, M. (2020). Machine learning-based approach for depression detection in twitter using content and activity features. *IEICE Transactions on Information and Systems*, 103(8), 1825-1832.
- Arora, P., & Aeri, B. T. (2019). Burden of antenatal depression and its risk factors in Indian settings: A systematic review. *Indian Journal of Medical Specialities*, 10(2), 55.
- Chakraborty, K., Avasthi, A., Kumar, S., & Grover, S. (2009). Attitudes and beliefs of patients of first episode depression towards antidepressants and their adherence to treatment. *Social psychiatry and psychiatric epidemiology*, 44(6), 482-488.
- Chowdhary, K. (2020). Natural language processing. *Fundamentals of artificial intelligence*, 603-649.
- Ciccone, G., Sultan, A., Laporte, L., Egyed-Zsigmond, E., Alhamzeh, A., & Granitzer, M. (2018). *Stacked gender prediction from tweet texts and images notebook for pan at CLEF 2018*. Paper presented at the CLEF 2018-Conference and Labs of the Evaluation.
- Deshpande, M., & Rao, V. (2017). *Depression detection using emotion artificial intelligence*. Paper presented at the 2017 international conference on intelligent sustainable systems (iciss).
- Dinesh, S., & MuniRaju, Y. (2021). Scalability of e-commerce in the COVID-19 era. *International Journal of Research-GRANTHAALAYAH*, 9(1), 123-128.
- Eack, S. M., Greeno, C. G., & Lee, B.-J. (2006). Limitations of the Patient Health Questionnaire in identifying anxiety and depression in community mental health: many cases are undetected. *Research on social work practice*, 16(6), 625-631.



- Hemalatha, I., Varma, G. S., & Govardhan, A. (2013). Sentiment analysis tool using machine learning algorithms. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, 2(2), 105-109.
- Hull, D. A. (1996). Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1), 70-84.
- Jaakkola, T. S., & Jordan, M. I. (1997). *A variational approach to Bayesian logistic regression models and their extensions*. Paper presented at the Sixth International Workshop on Artificial Intelligence and Statistics.
- Kaur, J., & Buttar, P. K. (2018). A systematic review on stopwords removal algorithms. *International Journal on Future Revolution in Computer Science & Communication Engineering*, 4(4), 207-210.
- Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1), 1-167.
- Maree, M. (2021). Semantics-based key concepts identification for documents indexing and retrieval on the web. *International Journal of Innovative Computing and Applications*, 12(1), 1-12.
- Marie, M., SaadAdeen, S., & Battat, M. (2020). Anxiety disorders and PTSD in Palestine: a literature review. *BMC psychiatry*, 20(1), 1-18.
- Maree, M., Kmail, A. B., & Belkhatir, M. (2019). Analysis and shortcomings of e-recruitment systems: towards a semantics-based approach addressing knowledge incompleteness and limited domain coverage. *Journal of the American Society for Information Science*, 45(6), 713-735.

- Mattsson, S., Olsson, E. M. G., Carlsson, M., & Johansson, B. B. K. (2019). Identification of anxiety and depression symptoms in patients with cancer: comparison between short and long web-based questionnaires. *Journal of Medical Internet Research*, 21(4), e11387.
- Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4), 1093-1113.
- Nguyen, Q. T., Nguyen, T. L., Luong, N. H., & Ngo, Q. H. (2020). *Fine-tuning bert for sentiment analysis of vietnamese reviews*. Paper presented at the 2020 7th NAFOSTED Conference on Information and Computer Science (NICS).
- O'dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., & Christensen, H. (2015). Detecting suicidality on Twitter. *Internet Interventions*, 2(2), 183-188.
- Orabi, A. H., Buddhitha, P., Orabi, M. H., & Inkpen, D. (2018). *Deep learning for depression detection of twitter users*. Paper presented at the Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic.
- Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*.
- Plisson, J., Lavrac, N., & Mladenic, D. (2004). *A rule based approach to word lemmatization*. Paper presented at the Proceedings of IS.
- Porter, M. F. (2001). Snowball: A language for stemming algorithms. In.
- Priya, A., Garg, S., & Tigga, N. P. (2020). Predicting anxiety, ~~depression~~ stress in modern life using machine learning algorithms. *Procedia computer science*, 167, 1258-1267.

- Sadeque, F., Xu, D., & Bethard, S. (2018). *Measuring the latency of depression detection in social media*. Paper presented at the Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining.
- Shrestha, A., Serra, E., & Spezzano, F. (2020). Multi-modal social and psycho-linguistic embedding via recurrent neural networks to identify depressed users in online forums. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 9(1), 1-11.
- Tay, A. K., Riley, A., Islam, R., Welton-Mitchell, C., Duchesne, B., Waters, V., . . . Elshazly, M. (2019). The culture, mental health and psychosocial wellbeing of Rohingya refugees: a systematic review. *Epidemiology and psychiatric sciences*, 28(5), 489-494.
- Yalamanchili, B., Kota, N. S., Abbaraju, M. S., Nadella, V. S. S., & Alluri, S. V. (2020). *Real-time acoustic based depression detection using machine learning techniques*. Paper presented at the 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE).
- Zaghouani, W. (2018). A Large-Scale Social Media Corpus for the Detection of Youth Depression (Project Note). *Procedia computer science*, 142, 347-351.
- Zucco, C., Calabrese, B., & Cannataro, M. (2017). *Sentiment analysis and affective computing for depression monitoring*. Paper presented at the 2017 IEEE international conference on bioinformatics and biomedicine (BIBM).

## الملخص باللغة العربية

### نظام للكشف عن الاكتئاب السريري مبني على معالجة اللغة الطبيعية

الاكتئاب السريري هو اضطراب عقلي شائع يتميز بالاكتئاب المزاجي وفقدان الاهتمام / المتعة. تشمل الأعراض الأخرى انخفاض الطاقة ، والشعور بالذنب أو تدني قيمة الذات ، واضطراب النوم أو الشهية ، وضعف التركيز. وفقًا لمنظمة الصحة العالمية ، يؤثر الاكتئاب السريري على أكثر من 75٪ من الناس في جميع أنحاء العالم. عادة ، يعتمد فحص الاكتئاب السريري على التقييم النفسي من خلال المقابلات وجهاً لوجه. هذه الأساليب التقليدية تستغرق وقتًا طويلاً وعرضة للخطأ. أظهرت الأبحاث السابقة أنه من الممكن التعرف على مرضى الاكتئاب الإكلينيكي من خلال تحليل منشورات وسائل التواصل الاجتماعي. ومع ذلك ، لم تكن تحليلات محتوى وسائل التواصل الاجتماعي مقسمة وفقًا للأعراض المختلفة للاكتئاب السريري ولا بناءً على استغلال الموارد الخارجية للدلالات الطبية. بالإضافة إلى ذلك ، يمكن أن تتميز حلول الفحص الحالية بطابع غير محكم الاقتران ، ومن ثم انخفاض مستوى صحة وموثوقية هذه الأدوات. على وجه الخصوص ، لا تعتمد صحة هذه الأساليب على البنية الرياضية للأعراض في بنية متعددة الأبعاد. وفقًا لذلك ، قمنا بتطوير نظام مقترن بقوة يجمع بين معالجة اللغة الطبيعية (NLP) وموارد المعرفة الطبية لمساعدة المتخصصين في الرعاية الصحية في فحص الاكتئاب السريري ؛ نشر أداة موثوقة وفعالة يمكنها المساعدة بشكل سلبي وتلقائي في تحديد الأشخاص الذين يعانون من أعراض الاكتئاب السريري بناءً على منشوراتهم على وسائل التواصل الاجتماعي. نظرًا للحالة المزمنة لخدمات وموارد الصحة العقلية، من المتوقع أن يعالج النظام المقترح الصلة السريرية المباشرة لأنه يدمج (Beck Depression Inventory II) (BDI-II) مع التحليل اللاحق لوسائل التواصل الاجتماعي باللغة الإنجليزية والعربية.

بالنسبة للغة الإنجليزية ، استخدمنا مجموعة بيانات eRisk 2020 الخاصة بمنتهى التقييم متعدد اللغات (CLEF) وهي مجموعة بيانات عالمية تتضمن منشورات على وسائل التواصل الاجتماعي للأشخاص المصابين بالاكتئاب بمستويات مختلفة ، ومنشورات الأشخاص الذين تظهر عليهم علامات المقامرة المرضية مع وجود علامات تدل على وجودهم. إيذاء النفس. هدفنا في هذا السياق هو معرفة ما إذا كانت منشورات وسائل التواصل الاجتماعي تعكس وجود أعراض الاكتئاب الإكلينيكي وكذلك شدتها. للقيام بذلك ، نستخدم موارد دلالية متعددة في مجال الطب النفسي لتعيين محتوى منشورات وسائل التواصل الاجتماعي إلى الأعراض المقابلة للاكتئاب السريري في BDI-II. لقد استخدمنا نتائج هذه الخطوة لتدريب النظام المقترح وتطوير إجراء فحص موثوق وفعال. كما تشير النتائج التي تم الحصول عليها ، فإن استخدام النموذج المقترح ودمجه مع تمثيلات التشفير ثنائي الاتجاه من نموذج المحولات (BERT) أدى إلى معدل دقة بنسبة 87 ٪ ، مع دمج النموذج المقترح مع المصنفات الأخرى ، مثل الانحدار اللوجستي ، و Naive Bayes ، أنتجت المصنفات XGBoost و Support Vector Machine (SVM) و Random Forest (RF) 55.8 ٪ و 50 ٪ و 47.4 ٪ و 57 ٪ و 49 ٪ على التوالي.

من ناحية أخرى ، بالنسبة للغة العربية ، استخدمنا مجموعة بيانات (أطلق عليها اسم مجموعة البيانات) لتطوير خط أنابيب يعتمد على البرمجة اللغوية العصبية (NLP) لاكتشاف الأشخاص المكتئبين بناءً على منشوراتهم على وسائل التواصل الاجتماعي المكتوبة باللغة العربية. لتقييم النموذج المطور ، استخدمنا العديد من المصنفات القائمة على التعلم الآلي باستخدام

مجموعة البيانات العربية وحققنا معدل دقة بنسبة 81٪ باستخدام SVM ودقة 70٪ باستخدام مصنفات الانحدار اللوجستي (LR) و NB و RF ، على التوالي.