eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Bachelor Thesis

Informatics Engineering Degree

Computation

Chronological detection of depression in social media threads by means of natural language processing

Asier Garin Aldezabal

Advisors
Arantza Casillas
Maite Oronoz

September 17, 2023

Acknowledgements

I would like to express my sincere appreciation and gratitude to my advisors for their invaluable contributions and support throughout this endeavour.

Special thanks to my tutors, including Alicia Pérez whose contribution has been crucial. Their dedication, expertise, and guidance have been instrumental in making this thesis possible.

To my parents and classmates for the unwavering support and encouragement have been a source of motivation and strength, enabling me to overcome challenges and reach new heights.

I am deeply thankful to everyone that has helped me for their unwavering commitment and exceptional contributions, and I am honoured to have had the opportunity to work alongside such remarkable individuals.

Abstract

Detecting depression in social media has become an increasingly important research area in recent years. With the widespread use of social media platforms, individuals at risk of suicide often express their thoughts and emotions online, providing an opportunity for early detection and intervention.

Artificial Intelligence and, particularly, Natural Language Processing open pathways towards the processing of massive amount of messages and the detection of depression traits and other risks related to mental health. Our main thesis question rests on the early prediction of depression detection in social media messages. We explore the accuracy gained by a system as more and more information (in terms of more social messages over time) from a user are available. Is the system becoming more and more accurate given subsequent information or is there a limit? How many messages do we need to train a simple model capable to attain an accuracy above a threshold? Do recent messages add much information to older ones? These research questions have arisen in our work.

A key cornerstone in artificial intelligence-based approaches rests, needless to say, on to the available data-sets. The data available bounds the ability of the system to gain knowledge. Thus, an important part of this work consists on an overview of the data-sets used to detect depression in social media, also mentioning various extra data-sets along the way. In our study we found that there are international challenges devoted to this task, among others, CLPsych.

We explore simple though efficient inference algorithms able to classify messages; next, we test the ability of the models to classify a user as with or without risk, just given social messages written by the user. In an attempt to put the focus on our main research question (i.e. assessing the impact of getting more and more information across time to gain accuracy in the task of message classification in the frame of early detection of depression signs) we opted for simple classifiers, that is, linear approaches, and left out of the scope exploring the behaviour of different classification approaches. Our experimental framework is developed using the practice data-set made available at CLPSych 2021. To make use of the data more intelligently, the chronological factor is added. Using a specific technique that progressively takes into account new data (chronologically) at each time, we can observe promising changes in the classification accuracy. These values might provide key ideas about the evolution of depression signs for detection. In other words, the results in a time-line might help to gain evidences that a user might be showing traces of or towards depression.

At the end, some comparisons and discussion are made regarding past research work related to this field, to do a critical analysis of the results.

Contents

List of Figures

List of Tables

List of Algorithms

CHAPTER 1

Introduction

In this chapter, we will briefly describe the thesis surroundings, in order to clear up the framework of the thesis.

In the ever-evolving landscape of information technology, the fusion of artificial intelligence and linguistics has given rise to a transformable field known as Natural Language Processing (NLP) [1]. Simply put, NLP stands as a testament to humanity's quest to bridge the communication chasm between humans and machines, enabling computers to comprehend, interpret, and generate human language. This thesis takes advantage of the NLP processes to achieve our goals related to depression detection.

As for many artificial intelligent fields, the need for initial data is essential. The input data for Natural Language Processing (NLP) consists of textual information in the form of human language. This language can be either written or spoken. NLP algorithms and models process this data to understand, analyze, and generate meaningful responses. The input data can vary in complexity and length, ranging from short sentences and paragraphs to lengthy documents or even conversations. Here are some examples of input data for NLP:

- Textual Documents: This includes articles, research papers, news stories, emails, legal contracts, and any written content.
- Social Media Posts: Tweets, Facebook posts, Instagram captions, and comments are examples of user-generated content that NLP can process for sentiment classification and prediction.
- Textual Data from Sensors or Devices: Textual data generated by devices, sensors, or IoT devices can be analyzed using NLP for insights.
- Medical Records and Reports: Medical texts, including patient records, research articles, and clinical notes, can be processed for information extraction or decision support.
- Translation Data: Pairs of text in different languages for translation tasks.

- Chatbot Interactions: Conversations between users and chatbots can be used for improving the responses and training of chatbot systems.

In this thesis we limit ourselves to the use of Social Media Posts. Apart from the text of the posts we can use the timestamp of each post to experiment with it. These type of data can be found in large data-sets or corpora, where the posts of some anonymous users are documented. The data is usually quite accessible but it can always bring up some issues. Here are some of them:

- Short Text Length: Social media posts often have character limits (e.g., Twitter's 280 characters). This brevity can lead to limited context and make it challenging to understand the intended meaning.
- Slang and Informal Language: Social media users often use slang, abbreviations, and informal language that might not be present in traditional dictionaries. Understanding these expressions requires specialized language models.
- Emojis and Emoticons: Emojis and emoticons are used extensively in social media to convey emotions or meanings. Apart from the challenge of interpreting their nuances and context, emojis are usually translated to text too (Unicode format, that way it can be included in the standard usage of UTF-8)[2].
- Spelling and Grammar Variability: Social media users might not adhere to traditional spelling and grammar rules. Misspellings, neologisms, and unconventional sentence structures are common.
- Hashtags and Mentions: Hashtags and mentions are essential in social media for categorization and conversation. Understanding their role and context requires specialized handling.

Most of these problems can be approached with Text Preprocessing, a crucial Natural Language Processing (NLP) step in preparing textual data for machine learning tasks. It involves various techniques to clean, format, and structure text data so that it can be effectively used by the model.

Fortunately, using text for NLP also has advantages. For example, in our case, there is information that we extract by processing social media posts:

- Sentiment Analysis: Determine the sentiment of a post—whether it is positive, negative, or neutral. In our case, we determine if the user has depression or not.
- Emotion Detection: Detect specific emotions expressed in posts, such as joy, anger, sadness, or surprise. In our work, depression is the emotion we detect and it is deeply related with the possibility of the user attempting suicide.
- Predictive Insights: Analyze historical data to make predictions about future, such as predicting if a user is going to have depression.

Apart from the information that we can acquire from the text itself, every post is usually embedded with:

- Timestamp: The date/time the post uploaded.
- The users identification number (anonymous).
- The posts identification number (anonymous).
- The posts text (title and/or content).
- (Optional) Users label if needed (in our case, depressed or not).

We will mainly focus on depression detection. We will also make use of the timestamps of the posts to approach some early predictive insights regarding depression.

To be more specific, the following is an example of the development of depression signs of a user by adding more text (current/newer posts) to the model. In other words, it is a visual representation of how confident the classifier is that the user shows traces of depression. The classifier does not change a lot at first, but during the last posts it gathers enough information to evaluate the user correctly, as *depressed*.

i	x_i	$w(depressed \vert x_1 \dots x_i)$	$\hat{y}(x_1 \dots x_i)$
1	Ion Care What Nobody Gotta Say Bout It Or Me!		
2	Just posted a photo		
3	Just posted a photo		
4	I Was Drunk Outta There Lol I Had Fun Tho!:)		
5	Lol This Africa Lady Carryn This Jug Of Juice On Her Head!	0.413	No
6	Ian On That Negative Side		
7	Me As Of Now And Today!:)		
8	New Pickup:)Goin To Get More.Next Week		
9	Just Finished Nae Hair!		
10	Can't Wake Up Bein Negative,I'm Not A Negative Person!	0.415	No
11	Ain't No Entertainment..Wea I Can Laugh		
12	Just posted a photo		
13	Just Got Out The Tub...!Bored Now		
14	Just Did My Hair:)		
15	Ian Know It Was This Late!	0.604	Yes

Table 1.1: example of the model evaluation, First 15 messages published by a 'Depressed' user. i: message cardinal; x: input text (we explored in batches of 5); $w(depressed|x_1 \dots x_i)$ confidence-score provided by the classifier for the messages seen up to the current input (i), that is, with the messages seen so far, how confident is the classifier that the user shows traces of depression (the bigger, the more confident); $\hat{y}(x)$: label estimated by the classifier (depression/control)

1.1 Motivation

The work aims to shed light on the potential of NLP in identifying and addressing mental health challenges, particularly in the context of early depression detection.

There are many other research papers, previous to this one. The method they use involves obtaining initial data from a corpus composed of any social media posts (Reddit, Twitter...), processing the data, employing Natural Language Processing (NLP) techniques and machine learning approaches to train the data and evaluate the results. Here are some examples which illustrate this kind of work [3][4]. An overview of the antecedents will be given in Chapter 2.

The method we are going to use is inspired on the mentioned one, making our work efficient and solid. Moreover, we found of interest to go beyond and tackle a gap found in those works, that is, to explore the early detection of a certain sentiment (depression).

Furthermore, this context joining NLP and mental health, there are also yearly celebrated competitions that promote early detection and related research. We devote to these initiatives a section afterwards, in section 2.1.

The thesis seeks to inspire hope and create awareness about the possibilities offered by NLP in proactively identifying signs of depression at an early stage. It delves into the author's journey, recounting the challenges faced, and the intriguing results over adversity. Through introspective details, the thesis highlights the crucial role that NLP can play in early intervention, potentially saving lives and mitigating the long-term effects of depression. Additionally, this type of NLP research improves access to mental health services and contributes to the creation of personalized and effective interventions.

In summary, the motivation behind a thesis centred on early depression detection by NLP is to contribute to the implementation of advanced technologies that can positively impact the lives of those affected by depression.

1.2 Goals and Objectives

We have organized our work in three different main goals and divide them into different sub-objectives.

As our work consists of depression detection using NLP techniques in social media posts, we can prioritize that as the first main goal. The second main goal is related to obtaining an early detection of depression with some confidence. The last main goal is directed to managing the thesis, including procedures on how to develop it.

1.2.1 First Main Goal: Framework Development

Design and implement a robust NLP framework for analyzing social media posts to identify signs of depression. The sub-objectives consists of the following:

1. Background Research
    - Gather and review existing knowledge in the field of early depression detection via NLP
2. Data Management
    a) Corpus Search

- Identify suitable data-sets or corpora for depression detection through text analysis.
- Focus on social media data sources for the research.
- Analyze one or more corpora to determine their suitability for depression detection.

b) Data Processing

- Retrieve and manage relevant information from the data-sets for analysis.
- Extend the analysis to another corpus. Perform the process in a different corpus in order to generalize the results.

3. Depression Detection

- Implement NLP techniques on selected data-sets.
- Define a way to identify individuals with depression for those that do not suffer that disorder.

1.2.2 Second Main Goal: Early Detection

Develop a reliable method for early detection of depression in individuals through the automatic analysis of social media content. The sub-objectives consists of the following:

1. Longitudinal Analysis

- Investigate the potential for early detection by analyzing data chronologically.
- Monitor changes in the results as newer posts are added to the model. Ultimately, exploring the possibility to determine a posts quantity threshold for identifying individuals with depression.

2. Results Presentation

- Effectively present experimental results using visual aids like plots and tables.

1.2.3 Third Main Goal: Project management

Organize and manage the work, follow and control the process. The sub-objectives consists of the following:

1. Thesis Management

- Properly manage the thesis, including resource allocation, documentation, and conclusion drawing.
- Brainstorm and refine thesis objectives and reachability for clarity and guidance.
- Create a comprehensive thesis document detailing methodology, findings, and conclusions.

2. Code Development

- Develop and document thesis code sequentially with input from advisors.

3. Iterative Development
    - Embrace an iterative work methodology to adapt to changing thesis needs and outcomes.
    - Maintain open communication with advisors for guidance and clarification throughout the thesis.

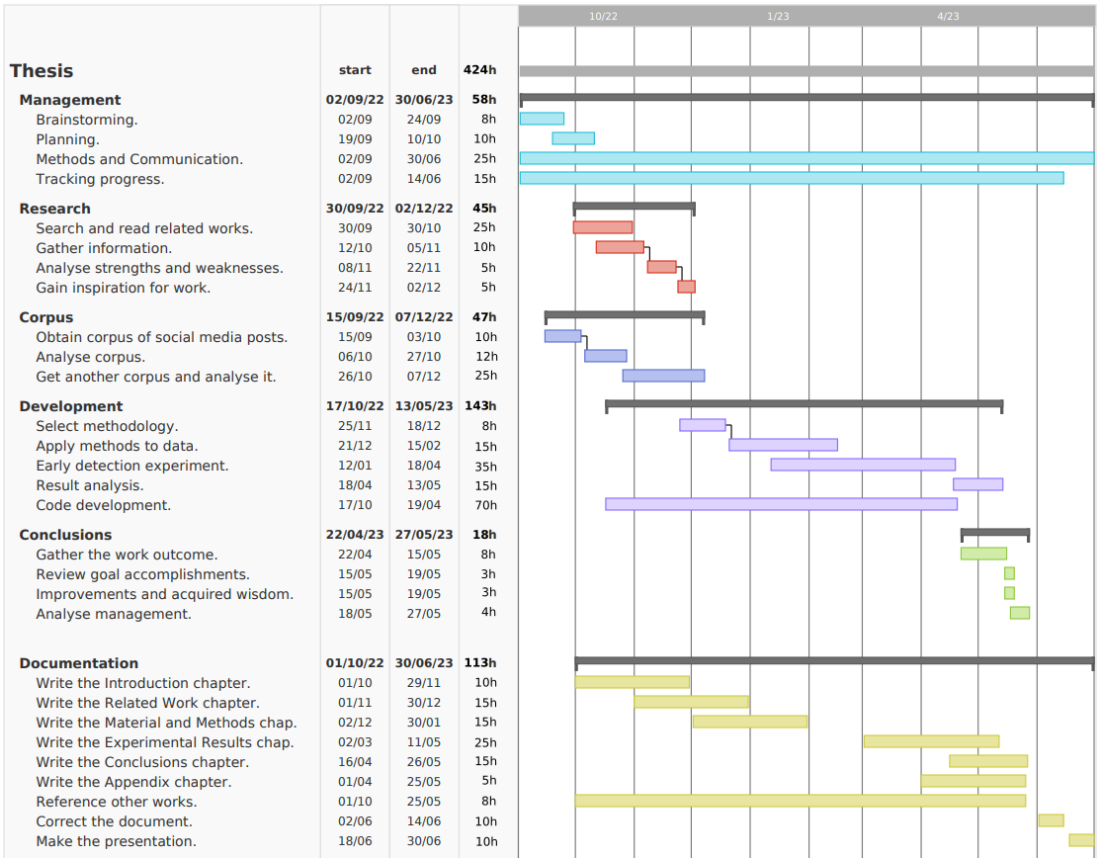
In this revised organization, the main objectives focus on the most critical aspect of our thesis, while the sub-objectives support and facilitate the achievement of this primary goals. Furthermore, in order to complete the proposed goals, we decompose the thesis into smaller work areas which include tasks organized in a hierarchical manner. More about this matter in the following section.

1.3 Project Management

We will start by mentioning the work areas. This breakdown allows for a clear understanding of the thesis's components and facilitates effective thesis management. Each work area can wrap up one or more goal and they are composed by tasks. These are the work areas:

- Management: This work area includes thesis management tasks such as initial brainstorming of objectives and reachability, structured work planning and visualization of working methods.
- Research: Before starting the development, it is necessary to have knowledge of the previous background in the area of early depression detection via NLP. The gathered related work inspires our methodology and development process. That is why this work area includes tasks for information gathering and instructive learning.
- Corpus: To be able to develop the thesis we need corpus, data-set or data-sets. This work area includes the search of the data-sets that can be useful for depression detection via text. The most suitable data-sets that contain text for this type of research are social media corpora, that contain posts of a bunch of users. This work area also involves analysing the data-sets.
- Development: Once we have chosen the data-sets that will best help us achieve our objectives, we will get to work. This work are includes methodology selection, using the Natural Language Processing framework for experiments, using the data chronologically for early detection and analyzing the results. Nevertheless, this work area includes the code programming work that is done in the background.
- Conclusions: Not only reflect the outcomes of our research, but also shed light on the broader implications and significance of our findings in the context of depression detection. They serve as a testament to the depth of knowledge and understanding we have gained throughout this immersive experience. This work area involves overall conclusions, acquired wisdom and possible enhancements/improvements of the thesis.
- Documentation: Alongside all the work, a thesis will be developed to document the work. Subsequently, the corresponding thesis presentation will be created for its defense.

The tasks compose the work areas and measure more precisely the time spent on them. The Gantt diagram is the ideal chart for our need, to appreciate a clear overview of the work timeline, tasks and spent hours we preview. The diagram can be observed in figure 1.1.



Thesis	start	end	424h
Management	**02/09/22**	**30/06/23**	**58h**
Brainstorming.	02/09	24/09	8h
Planning.	19/09	10/10	10h
Methods and Communication.	02/09	30/06	25h
Tracking progress.	02/09	14/06	15h
Research	**30/09/22**	**02/12/22**	**45h**
Search and read related works.	30/09	30/10	25h
Gather information.	12/10	05/11	10h
Analyse strengths and weaknesses.	08/11	22/11	5h
Gain inspiration for work.	24/11	02/12	5h
Corpus	**15/09/22**	**07/12/22**	**47h**
Obtain corpus of social media posts.	15/09	03/10	10h
Analyse corpus.	06/10	27/10	12h
Get another corpus and analyse it.	26/10	07/12	25h
Development	**17/10/22**	**13/05/23**	**143h**
Select methodology.	25/11	18/12	8h
Apply methods to data.	21/12	15/02	15h
Early detection experiment.	12/01	18/04	35h
Result analysis.	18/04	13/05	15h
Code development.	17/10	19/04	70h
Conclusions	**22/04/23**	**27/05/23**	**18h**
Gather the work outcome.	22/04	15/05	8h
Review goal accomplishments.	15/05	19/05	3h
Improvements and acquired wisdom.	15/05	19/05	3h
Analyse management.	18/05	27/05	4h
Documentation	**01/10/22**	**30/06/23**	**113h**
Write the Introduction chapter.	01/10	29/11	10h
Write the Related Work chapter.	01/11	30/12	15h
Write the Material and Methods chap.	02/12	30/01	15h
Write the Experimental Results chap.	02/03	11/05	25h
Write the Conclusions chapter.	16/04	26/05	15h
Write the Appendix chapter.	01/04	25/05	5h
Reference other works.	01/10	25/05	8h
Correct the document.	02/06	14/06	10h
Make the presentation.	18/06	30/06	10h

Figure 1.1: Gantt diagram of our work estimation, throughout an eight month period of time. In the first column the tasks and respective work area appear. In the second and third columns the beginning and the end of each task, respectively. Lastly, in the fourth column, the estimation of the working hours for each task.

1.4 Document Structure

- Chapter 1. Introduction

  The thesis starts with this chapter, the introduction. It starts by mentioning, explaining the background and inspiration of this thesis, including various conventions. It follows by explaining the motivation behind this thesis, then mentioning the goals and ends providing this structure of the thesis to give an initial perspective of it. The sections that compose this chapter are the following:

  1. Motivation: We explain why our thesis is essential, highlighting the significance and relevance of our work in the broader context of our field. Additionally, we mention the antecedent research that inspire this work.

2. Goals and Objectives: Outlines the specific outcomes and achievements we intend to accomplish through our thesis.
3. Project Management: Work breakdown in work areas and task. In addition, a Gantt diagram with the estimated timeline and ours for each task.
4. Document Structure: An initial overview of the document chapters and section.

- Chapter 2. Related Work

  We provide a comprehensive overview of existing events and research that is relevant to our work. This section serves several purposes, including contextualizing our work within the broader field, demonstrating our understanding of prior research, and identifying gaps or opportunities that our research addresses. The sections that compose this chapter are the following:

  1. Background: Information about the antecedent research, including events and research.
  2. Strengths and Weaknesses: Evaluation of the prior work of other researchers, contemplating the reinforcements and considerations.

- Chapter 3. Materials and Methods

  We provide a detailed description of the materials used in our work and the methods employed to conduct our thesis. This section is comprehensive and transparent to enable other researchers to replicate our work. The sections that compose this chapter are the following:

  1. Material: List, description, specifics and analysis of all the materials we used in our work.
  2. Methodology: A brief description of the methods used in the work is provided, accompanied by an overview of the Natural Language Processing (NLP) techniques and its underlying mathematical foundations. The section concludes by explaining the evaluation metrics employed in the study.

- Chapter 4. Experimental Results

  This chapter explores the step-by-step development process of the early depression detection thesis using NLP. It provides a comprehensive overview of the methodologies, algorithms, and data sources employed in the thesis. The author discusses the challenges encountered during the development phase and the innovative strategies employed to overcome them. The sections that compose this chapter are the following:

  1. CLPSych2021 Practice data-set results: We, initially, train and test the model with this data-set. We show the results obtained from testing the data-set, the practice data-set.
  2. CLPSych2021 Reddit data-set results: We repeat the same train/test process. We show the results obtained from testing data-set, the Reddit data-set. A much larger and significant data-set than the other one. Afterwards, we compare the results obtained from the two data-sets.

3. Data examples: Visual examples of the format in which we obtain the data and how to interpret the results obtained from the model.
4. Error Analysis: We comment the behaviour of the model and illustrate relevant observations, if any. The analysis is mainly done in the CLPSych2021 Reddit data-set, the CLPSych2021 Practice data-set results are not meaningful.
5. Discussion: We review the results obtained in our work. We compare the obtained results with the results obtained in other related works.

- Chapter 5. Conclusions

  In this concluding chapter, the author reflects on the entire thesis and its impact. We summarize the main achievements, compare the obtained results from the data-sets, discuss the occurred adversities and summarize the gained acknowledgement. The chapter also includes a compelling number of possible enhancements to improve the thesis and obtain further better results. The chapter concludes by proposing a bunch of changes that can be applied to the thesis in order to get a different perspective and improve it in an alternative way. The sections that compose this chapter are the following:

  1. Conclusion: Recap of the goals, analysis of achievements.
  2. Project Management: Comparison between the forecast and the actual reality of the time spent in each task. We also mention the main deviations of the work.
  3. Acquired knowledge: Recount of the valuable knowledge we gained from this work.
  4. Enhancements and future improvements: Ways to follow this work, improvements.

- Chapter Appendix

  The thesis includes an appendix that provides supplementary materials that the reader might be interested in. The sections that compose this chapter are the following:

  1. Useful Links: List of links that facilitate the work of the ones that are researching in the same path.
  2. Related National Challenges/Competitions: A set of additional national challenges related with detection of emotions, computational linguistics and NLP.
  3. Additional Corpus: Corpus that we came across during the search of the two data-sets we use.

- Final Chapter Bibliography This chapter serves as a valuable reference for readers interested in the technical aspects of the thesis and encourages further exploration.

CHAPTER 2

Related Work

Before getting down to work, this chapter serves as a critical foundation for our research, allowing us to position our study within the larger academic context. The chapter delves into the landscape of previous studies, theories, and advancements in depression/anxiety detection. By exploring previous works, we aim to understand the evolving nature of the field and to identify gaps that our research can address. Through this comprehensive survey, we strive to build upon established knowledge, integrate key theories, and acknowledge the contributions of our predecessors.

2.1 Background

Firstly, we present the predecessors and inspiration of this thesis. We set the base of our work with them, the methodology and material are thoroughly inspired by them.

2.1.1 Interdisciplinary events

For starters, there are interdisciplinary events that bring together experts and researchers from both the computational linguistics and clinical psychology domains. These workshops provide a platform for discussing the application of Natural Language Processing (NLP) and computational techniques in understanding, diagnosing, and treating mental health and psychological conditions.

These workshops focus on bridging the gap between the fields of linguistics and psychology by showcasing the latest advancements in using NLP and machine learning to analyze textual data, such as social media posts, chat logs, and clinical narratives. These workshops aim to foster collaboration between computational linguists and clinical psychologists, encouraging the development of innovative approaches for detecting mental health issues, predicting psychological states, and offering personalized interventions.
In our work, we will use some of the data-sets used in this events, more precisely, the ones of CLPsych 2021. It is helpful to mention that one of the data-sets we will use is the Version 2 of the data-set, currently available, it includes the training and test data from the 2019 CLPsych Shared Task.

CLPsych 2019: Also known as the Sixth Workshop on Computational Linguistics and Clinical Psychology [5]. It was organised in conjunction with the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), one of the premier conferences in the field of Natural Language Processing (NLP) and computational linguistics. The main focus of CLPsych 2019 was the intersection of computational linguistics, natural language processing, and clinical psychology, with a particular emphasis on using computational methods to analyse and understand mental health-related text data. A Shared Task was held that focused on predicting individuals' suicide risk from de-identified, public Reddit data. Suicide risk is directly related with depression, we consider that anyone that has attempted suicide is suffering from depression. In consequence, as part of our work is to detect depression, the Reddit data could be useful. One of the proposed tasks, the most relevant for us, involved predicting level of risk for users posting to the r/SuicideWatch Subreddit based on their SuicideWatch posts. More information of these events, such as CLPysch or eRisk is mentioned in the appendix 5.4.

CLPsych 2021: Also known as the Seventh Workshop on Computational Linguistics and Clinical Psychology [6] follows the same principles as the previously explained workshop. They differ in the presented Shared Task, this year it consisted of using sensitive data in a secure data enclave. Bringing researchers to the data rather than sending the data out to researchers. Participating teams received access to Twitter posts donated for research using OurDataHelps.org platform, including data from users with and without suicide attempts. All the work was done with the data-set entirely within a secure computational environment provided by NORC at the University of Chicago. In addition, they also gave the option to use the *University of Maryland Reddit Suicidality Dataset, Version 2*; which contained it includes the training and test data from the 2019 CLPsych Shared Task.

2.1.2 Antecedent articles/research

Besides the mentioned events there are singular research papers that experts from all over the world release to the public. Some of those research papers are directly correlated with this work, it could be said that they are antecedents of this works. That is, we were inspired by prior pioneering work in this domain, our work extends the boundaries of the prior work and uses the most fundamental parts. To give a notion of those prior works, we will mention a few of them, afterwards remarking the overall strengths and weaknesses.

Measuring the Latency of Depression Detection in Social Media [7]
The article provides information about the implementation of the RMSProp optimization algorithm for training models, the features used in the models, and the different approaches applied to predict depression status from user posts. The models incorporated count-based word features, depression word features, and other features such as posts per day and shared interactions. The document also mentions the use of ensemble classifiers, sequential models and non-sequential models.

Detection of Depression-Related Posts in Reddit Social Media Forum [3]
The document discusses the use of linguistic analysis and machine learning techniques for depression detection. It mentions the use of the LIWC2015 dictionary and topic modeling to extract lexical-syntactic features and hidden topics related to anxiety and depression from textual data. Various studies are referenced that explore the use of different features and classifiers for depression identification. The document also introduces the Reddit dataset as

a source of data for the study. The results show that combining features such as Linguistic Inquiry and Word Count (LIWC), Latent Dirichlet allocation (LDA), and bi-grams with machine learning classifiers can improve the accuracy of depression detection. A Multilayer Perceptron (MLP) classifier performs the best, achieving 0.91 accuracy and 0.93 F1 score.

Comparing emotion feature extraction approaches for predicting depression and anxiety from CLPsych 2022 [8]
Three feature extraction approaches were used in the study: BERT-based models, Linguistic Inquiry and Word Count (LIWC) 2015, and GoEmotions. These approaches extract emotions such as anger, sadness, positive emotion, and negative emotion. The variance in Patient Health Questionnaire (PHQ-9) scores explained by these variables was similar across the three approaches, with Linguistic Inquiry and Word Count (LIWC) explaining slightly more variance overall. However, for specific emotions like anger and sadness, other variables performed better.

Here are some more, this ones differ a bit as they are not directly connected with our type of work but they have many similarities that we used as inspiration.

Objective Assessment of Social Skills Using Automated Language Analysis for Identification of Schizophrenia and Bipolar Disorder [9]
The document focuses on the analysis of language features to predict Social Skills Performance Assessment (SSPA) performance and classify individuals with schizophrenia, schizoaffective disorder, and bipolar disorder. It considers semantic coherence measures, linguistic complexity measures, and a comprehensive set of language features. The results show the potential of these features in differentiating between clinical and healthy control groups and identifying specific clinical populations.

Multitask Learning for Mental Health Conditions with Limited Social Media Data [10]
Automated monitoring and risk assessment of patients' language have the potential to overcome the logistic and time constraints associated with traditional assessment methods for mental health conditions. Language carries implicit information about the author, which has been exploited in Natural Language Processing (NLP) to predict author characteristics and mental health conditions. Existing research indicates that incorporating demographic attributes and learning multiple auxiliary tasks can improve prediction performance. Multitask learning (MTL) models that predict multiple mental health conditions jointly show significant improvements over baselines and single-task models, particularly for conditions with limited data.

Multi-Task Learning for Mental Health using Social Media Text [11]
The document discusses the use of multi-task learning (MTL) in predicting mental health conditions based on social media text. The experiments compare the performance of logistic regression models, single-task feed-forward models, and multi-task feed-forward models. The results show that MTL can improve the prediction of mental conditions by leveraging commonalities and differences between tasks. The document also mentions the importance of feature representation and initialization in neural models.

2.2 Strengths & Weaknesses

As said, we will mention the strengths and weaknesses of some of the antecedent works, we will not specify which work has which strengths or weaknesses. Keep in mind that this

will help in the development of our thesis as we will exploit the weaknesses and include the strengths. Overall, we consider the weaknesses to be:

- **Incomplete evaluation metrics**: The work does not provide a comprehensive analysis of its limitations or how it compares to other evaluation metrics.
- **Limited information on specific studies**: The document mentions various studies and approaches related to text classification and depression detection, but it does not provide detailed information about these studies. This lack of specific details makes it difficult to assess the validity and reliability of the findings.
- **Lack of context**: The document does not provide a clear context or background information about the research problem or the data-set used. This makes it challenging to understand the significance and relevance of the findings presented.
- **Incomplete methodology description**: While the document mentions the use of certain techniques and features, it does not provide a comprehensive description of the methodology followed. This lack of detail makes it difficult to replicate or evaluate the study.
- **Limited discussion of results**: The document briefly mentions the performance and accuracy of different features and classifiers, but it does not provide a thorough analysis or interpretation of the results. This limits the understanding of the effectiveness of the proposed approaches.
- **Lack of comparison with existing literature**: The document does not discuss or compare the findings with existing literature on depression detection or text classification. This omission makes it challenging to assess the novelty or contribution of the study.
- **Limited perspective**: The method used in the study focuses only on the different models, metrics or features. There are other possibilities to explore, such as, data manipulation, chronological limitation regarding data usage...

And the strengths that we will consider in our work:

- **Training methodology**: The document explains the processes of using the data optimally to train the model.
- **Data collection**: The document describes the process of collecting texts for both depression language and general language. It mentions the sources of the texts and the number of posts collected for each category.
- **Acknowledgment of limitations**: The document acknowledges the possibility of having non-depressed individuals in the depression group or vice versa.
- **References**: The document includes a list of references at the end, which indicates that the information provided is supported by previous research and publications.

- **Use of machine learning techniques**: The document explores the use of machine learning techniques, such as Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR), to improve the accuracy of depression identification. This demonstrates the application of advanced algorithms in the field of mental health research.
- **Coherent results**: The document uses the pertinent metrics that are suited best for the specific task. That way it provides coherent results with the adequate metrics.
- **Data security**: The document mentions that the data used in the study were de-identified and stored on a secure server with limited access, ensuring data privacy and security.
- **Detailed methodology**: The document provides details about the methods used, including technical classifiers, enhancing the transparency and productivity of the study.

CHAPTER 3

Materials and Methods

In the pursuit of scientific inquiry, the systematic delineation of materials, methodologies, and procedures is essential to ensure the reliability and reproducibility of research outcomes. This chapter provides an in-depth exposition of the resources and techniques harnessed to realize the objectives of this work. By detailing the data-sets and methods employed, we not only establish a foundation for replication but also lay bare the framework through which our research inquiries were pursued.

3.1 Material

In this section we describe the data-sets and explain their origin, which is their composition, which are the labels, what type of information do they provide... All the needed characteristics to understand the data-sets. In each data-set there should be a README file that clarifies any possible doubt about the use, structure and size of the data-set. We will work with two different data-sets from CLPSych 2021, the first one smaller than the other. Information about additional corpora, apart from the ones below, can be found in the appendix 5.4.

3.1.1 CLPSych 2021

The twitter data-set with which we have performed our experiments. This data-set is the practice data-set that was provided in CLPSych 2021 [12]. They provided the practice data-set to help build the system outside the enclave (a network that's separated from the rest of the network and governed by granular security policy). This practice data-set is based on a modified version of *swcwang/depression-detection* [13]. The task is to identify users who have tweeted with a #depression (or similar) hashtag. The way of using the Practice Data-set is found here [14].

Note that although they performed spot checks to make sure this data-set seems reasonable, the practice data-set has not been validated by the community, so we have approached the results with scepticism.

3.1.1.1 Annotations

Data-sets like this one label the users 'SUICIDE' or 'CONTROL'. If a user has attempted suicide or not, respectively.

We have taken the liberty to relate suicide to depression. In other words, we have considered that if a user is suicidal, he/she is also suffering from depression; We have changed the label 'SUICIDE' by 'DEPRESSION'. The idea is supported by evidence such as:

> [es][original] "*¿La depresión aumenta el riesgo de suicidio? Aunque la mayoría de las personas que tienen depresión no se suicidan, el padecer depresión aumenta el riesgo de suicidio, sobre todo si ésta es grave. Así, cerca del 60-90% de las personas que se suicidan tienen síntomas de depresión.*"
>
> [en] "*Does depression increase the risk of suicide? Although most people who have depression do not commit suicide, having depression increases the risk of suicide, especially if it is serious. Thus, about 60-90% of people who commit suicide have symptoms of depression.*"

The quotes above are from the book [15]. Which is a academically supported book about suicide in general.

3.1.1.2 Details - Numbers

The data are provided in JSON-lines files (one for train and one for test), where each line represents a single user and their tweets. The format is as follows:

```
{
    "id": str, # anonymized user ID- used for submission
    "has_attempt": bool, # variable used in previous versions as "label"
    "date_of_attempts": str, # the known date of attempt or empty string if no attempt
    "label": bool, # true for depression hashtag, false for control
    "tweets": [
        {
            "id": str,  # unique id for each tweet
            "text": str, # text that the tweet contains
            "created_at": str # date of when the tweet was written
        }
    ]
}
```

The CLPSych 2021 organization explains the format and gives a brief description of the data structure. To get a better understanding of each parameter, we will explain them with a bit more detail.

- *id*
  It is the parameter used to identify the user. Each user has a different ID number in order to keep the user anonymous. The ID number is composed by 19 numbers.
- *has_attempt*
  The variable that was used in the previous versions of the data to know if someone had attempted suicide or not. Thus, has depression or not, respectively. Now appears as **null** in each user because it has migrated to the *label* parameter.

- *date_of_attempts*
  Supposed to be the date of the suicide attempt in case if the user did attempt it. But as mentioned, the *has_attempt* parameter is not used anymore.

- *label*
  It is the parameter that determines if a user has attempted suicide or not. It has the value *true* if the user has attempted and *false* if not. If the value is *true* we will assign DEPRESSION hashtag to the user and if the value is *false* the CONTROL hashtag. This is also explained in the *Annotations* sub-section above. So for now on, in order to achieve a better understanding, we will use DEPRESSION/CONTROL concepts when we talk about the users' classification.

- *tweets*
  List of the tweets of a user, each user has a different amount of tweets.

  - *id*
    The ID number of the tweet, each tweet ID, is different. The ID of the tweet is composed by 19 numbers as the user ID is.
  - *text*
    The text written on the tweet
  - *created_at*
    The date that the tweet was written.

Each user has a various amount of information, as well as the tweets. With the intention of clarifying how an instance of the data-set looks, here are some examples:

- For user **781790505776676864**


```
{"id": "781790505776676864",
"has_attempt": null,
"date_of_attempts": null,
"label": false,
"tweets": [{
    "id": "1340351011244953600",
    "created_at": "2020-12-19 17:39:18 UTC",
    "text": "@mwalimu001 the wrath of ozil is imaging"},
    {"id": "1322945759529082881",
    "created_at": "2020-11-01 16:57:02 UTC",
    "text": "thermos party is wonderful for arsenal bravo @Fadhilow "},
    {"id": "1317019690108112896",
    "created_at": "2020-10-16 08:28:57 UTC",
    "text": "what do we call this type of protocol @mwalimu001 @Mediphaz"}]
}
```


- For user **1148236784490438656**


```
{"id": "1148236784490438656",
"has_attempt": null,
"date_of_attempts": null,
"label": true,
"tweets": [{
```

```
            "id": "1341748346772467712",
            "created_at": "2020-12-23 14:11:49 UTC",
        "text": "@qu_qian The quote says that life is a present as in life is a \"gift\""},
            {"id": "1339718816956354563",
            "created_at": "2020-12-17 23:47:11 UTC",
            "text": "@qu_qian It's a nightmare for many"}]
}
```


To sum up all the explanation of the data structures, we will mention the most trivial parts that are in fact related with each other:

- The *has_attempt* parameter is never used but it has the same function as the *label* parameter. In reality, it has no use and its value is always **null**, as well as the parameter *date_of_attempts*. Originally, it was used to know the dates which the user had attempted suicide.
- We use the *label* parameter to know the classification of the user:
  - **Control** classified user, when a user has *false* in the label parameter. It means that the user has not attempted suicide.
  - **Depression** classified user, when a user has *true* in the label parameter. It means that the user has attempted suicide.

As we said, the terminology Control/Depression is going to be used from now on to simplify the understanding of the user classification.

The current data-set is the *practice_database* given in the CLPSych 2021 event. It can be acquired by going to GitHub to the *clpsych2021-shared-task* project [14]. In the GitHub there is a README file explaining the procedure to get the data or the tweets of each user. It goes as the following: We are given a data file that is in a dehydrated form, in that data file we only have the IDs of the users and tweets. Knowing that, with a tool called Twint and a Python script we *hydrate* the files, which in other words means we obtain the text of each tweet. Until now everything seems logical, but there is a problem.
Due to some reasons, such as, errors with extracting the tweets from modern web browsers (with the Twint tool) or some tweets just being deleted, a large part of the dehydrated file is lost in the *hydration* process. Not only that, as time goes by, when we try to *hydrate* the file, the number of tweets we get decreases. The largest number of tweets we got from the *hydrating* process was back in June of 2022. Although it was the biggest number of tweets we got, the loss of hydrating the initial file is bigger than expected, see tables 3.1 and 3.2.

As we can see in the tables 3.1 and 3.2, the number of tweets we got was extremely lower than the ones that appeared on the *dehydrated* file. Only the 0.6% of the tweets were retrieved in the General training set and only 0.2% in the Test set. As said, this is due to Twint module problems in adapting to the fast changes in Twitter and web-browsers. This loss of tweets causes that some users have no tweets within.

Regarding the data distribution, there are two main sets of the data-set, the ***General Train*** set and the ***Testing*** set. We have decided that the ***General Training*** set will be partitioned in a stratified way into two sub-sets, one for *training (**train**)* and the other for

General Training Set			
	Dehydrated File	Hydrated File	Loss
Total number of users	1262	1198	64
Users with tweets	1262	177	1085
Total number of tweets	800015	5077	794938
Tweets per user mean ± stdev	634±297	29±26	-
Users of Control	631	80	551
Users of Depression	631	97	534

Table 3.1: The lost information in the process of *hydratation* in the ***General Training*** Set.

Testing Set			
	Dehydrated File	Hydrated File	Loss
Total number of users	66	61	5
Users with tweets	66	8	58
Total number of tweets	40179	92	40087
Tweets per user mean ± stdev	609±323	27±22	-
Users of Control	33	6	27
Users of Depression	33	2	31

Table 3.2: The lost information in the process of hydratation in the ***Testing*** set

*development/testing (**dev**).* The partition, composed by 20% of the original training data assigned to ***dev*** and 80% assigned to *train*. The number of users of each class in the sub-sets are proportional to the one in the ***General Training*** set (the original set), that is, we have divided the data from the ***General Training*** set in a stratified way.

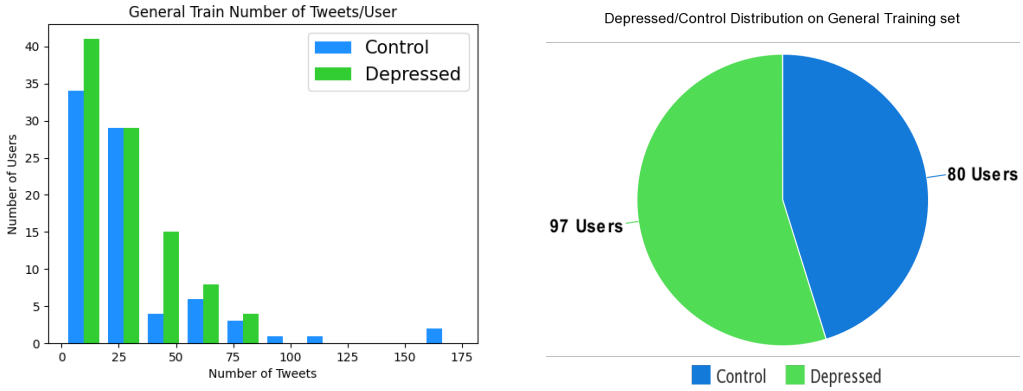
Although the two of them come from the training data, dev will be used for testing and will take that role in order to better train the model before testing it. More information about each of the sets and partitions follows.

- General Train Data-set
  We can observe in table 3.3 that the data is composed by 13373 different words. We have 5093 tweets in total and 177 users that contain tweets, as we previously saw in the table 3.1 there are some users that do not contain tweets. The number of tweets of the users is between [0-60] as it is shown in 3.1a histogram.
  Regarding the distribution of Control/Depressed users is mainly even. The number of tweets of each type of user is practically the same, as the median stands near 20 tweets. With a slight difference that Control users have a bigger standard variation. We can observe this in the 3.2 box-plot.

- Training Partition
  In this first data partition extracted from the General Train Set we can observe in table 3.4 that we have 4109 tweets in total and 141 users that contain tweets. The number of tweets of a user is between [0-60] like in the General Training set as it is shown in 3.3a histogram.
  Regarding the distribution of Control/Depressed users is mainly even. The number of tweets

General Training Set	
Total number of users	177
Total number of tweets	5077
Average ± Standard Deviation of the tweets of a user	29±26
Average ± Standard Deviation of the words on a tweet	8±6
Total number of words	42074
Size of the dictionary	13378
Users labelled as CONTROL	80
Users labelled as DEPRESSION	97

Table 3.3: ***General Training*** set instance details



(a) Histogram: number of users (Y) who posted a given number of tweets (X).

(b) Pie chart of users in set.

Figure 3.1: Measurements on the ***General Training*** set.

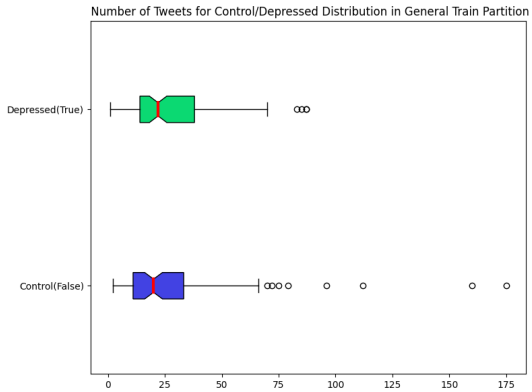


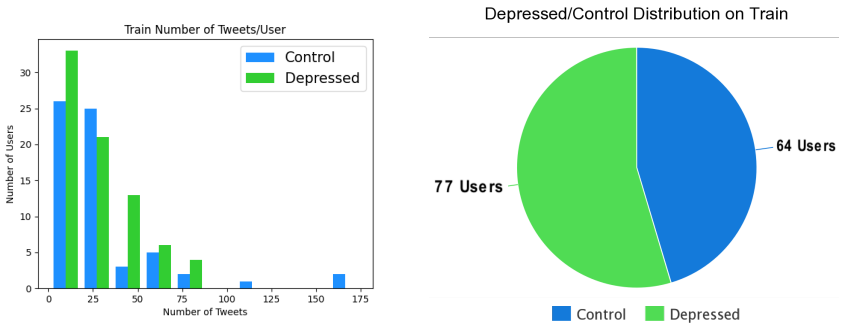
Figure 3.2: Box-Plot: distribution of number of tweets per-class on the ***General Training*** set.

of each user type is practically the same, with a moderate distinct of Control users having 20 median and the Depressed ones 25. Of course, the Control users have a bigger standard variation. We can observe this in the 3.4 box-plot.

- **Dev** Partition
  We can observe in table 3.5 that we have 981 tweets in total and 36 users that contain tweets.

Training Partition	
Total number of users	141
Total number of tweets	4096
Total number of words	34617
Average ± Standard Deviation of the tweets of a user	29±27
Average ± Standard Deviation of the words on a tweet	8.5±6.5
Users labelled as CONTROL	64
Users labelled as DEPRESSION	77

Table 3.4: Training partition instance details



(a) Histogram: number of users (Y) who posted a given number of tweets (X). **(b)** Pie chart of users in training partition.

Figure 3.3: Measurements on the ***Training*** partition.

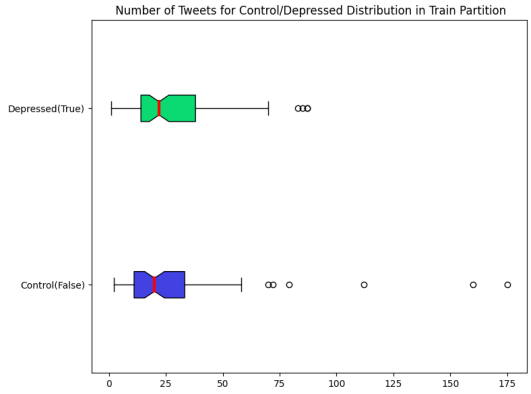
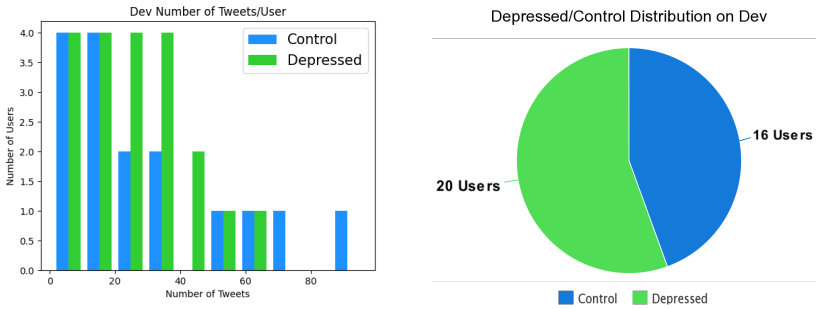


Figure 3.4: Box-Plot: distribution of number of tweets per-class on the ***Training*** partition.

The number of tweets of a user is between [0-50] similar to the General Training set as it is shown in 3.5a histogram.
Regarding the distribution of Control/Depressed users is mainly even. The number of tweets of each user type is practically the same, with a slight difference of Depressed ones having the median a bit higher. Of course, the Control users have a bigger standard variation. We can observe this in the 3.6 box-plot.

Dev Partition	
Total number of users	36
Total number of tweets	981
Total number of words	7457
Average ± Standard Deviation of the tweets of a user	27±22
Average ± Standard Deviation of the words on a tweet	7.5±6
Users labelled as CONTROL	16
Users labelled as DEPRESSION	20

Table 3.5: ***Dev*** partition instance details



(a) Histogram: number of users (Y) who posted a given number of tweets (X).

(b) Pie chart of users in ***Dev*** partition.

Figure 3.5: Measurements on the ***Dev*** partition.

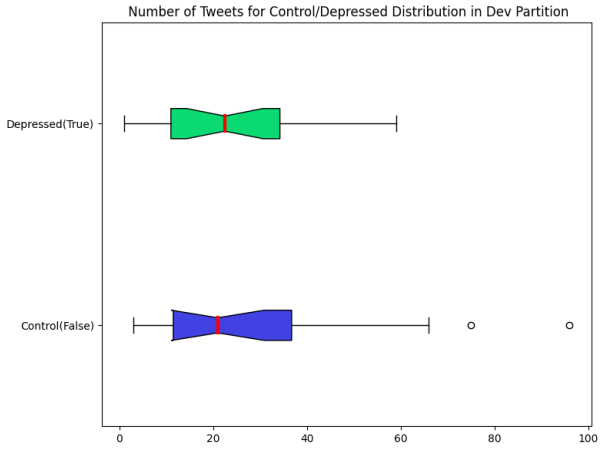


Figure 3.6: Box-Plot: distribution of number of tweets per-class on the ***Dev*** partition.

- **Test** set
  Lastly, many of the ***Testing*** set, the original one (not the one we extracted from the ***General Training*** set), tweets are lost in the *hydratation* process, so we will not use it. We can not talk much about it, the table 3.6 shows the necessary information. At least, the users of the set are labelled.

Test Data-set	
Total number of users	61
Total number of tweets	92
Total number of words	846
Average words on a tweet	9.2
Number of users without tweets	53
Number of users with one or more tweets	8
Users labelled as CONTROL	6
Users labelled as DEPRESSION	2

Table 3.6: Testing set instance details

3.1.2 Reddit Suicidality Dataset, Version 2

The University of Maryland Reddit Suicidality Dataset [16] was constructed using data from Reddit, an online site for anonymous discussion on a wide variety of topics, in order to facilitate research on suicidality and suicide prevention. The data-set was derived from the 2015 Full Reddit Submission Corpus [17], using posts in the r/SuicideWatch subreddit to identify (anonymous) users who might represent positive instances of suicidality.

They introduced Version 1 of the data-set in [18]. As reported there, annotation of users in this data-set by experts for level of suicide risk (on a four-point scale of no risk, low, moderate, and severe risk) yielded what is, to our knowledge, the first demonstration of reliability in risk assessment by clinicians based on social media postings. The paper also introduces and demonstrates the value of a new, detailed rubric for assessing suicide risk, compares crowdsourced with expert performance, and presented baseline predictive modelling experiments using the new data-set.

Subsequently, they updated the data-set for the shared task on predicting degree of suicide risk from Reddit Posts, run as part of the 2019 Computational Linguistics and Clinical Psychology Workshop (CLPsych 2019) held at the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), summed in [19]. Updates included adding automatic de-identification of post titles and bodies, as well as the definition of a standard training/test split to be used during the shared task in order to facilitate head-to-head comparisons of system performance. They also filtered out some posts from the Version 1 data-set based on encoding issues.

3.1.2.1 Details - Numbers

The data-set is accompanied by documentation about its format. Briefly, it contains one sub-directory with data pertaining to 11,129 users who posted on SuicideWatch, and another for 11,129 users who did not. For each user, we have full longitudinal data from the 2015 Full Reddit Submission Corpus, for each post:

- The post ID
- Anonymous user ID
- Timestamp
- Subreddit

- De-identified post title
- De-identified post body

For example a line in the data-set (one post):

```
2j7i1w,22002,1413286983,depression,Do you lost motivation to eat?,Question in
topic... i dont know what is wrong with me... i dont care about hunger anymore...
```

Although, the currently available Version 2 of the data-set includes the Training and Test sets data from the 2019 CLPsych shared task (621 users who posted on SuicideWatch and 621 who did not) with consensus annotations based on crowdsourcing plus the expert-annotated data (245 users who posted on SuicideWatch, paired with 245 control users who did not) which was not used in the shared task. We have used the data from crowdsourcing. In tables 3.7 and 3.8 show the details or information of the users and posts for the training and test partitions provided by the CLPsych 2021 organization.

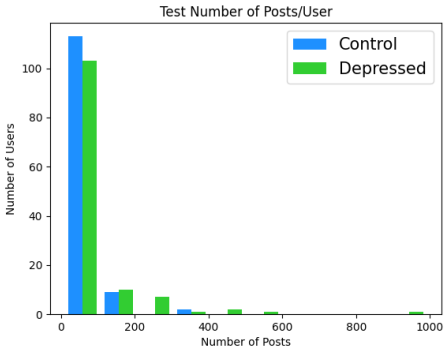
Training Set	
Total number of users	993
Total number of posts	56022
Average ± Standard Deviation of the post of a user	56,417±157,392
Average ± Standard Deviation of the words on a post	24,73±52,17
Total number of words	1385160
Size of the dictionary	70472
Users labelled as CONTROL	497
Users labelled as DEPRESSION	496

Table 3.7: Training set instance details.

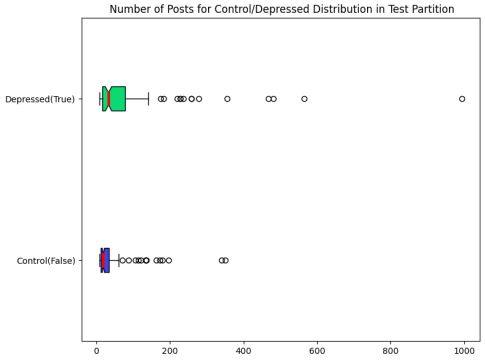
Test Set	
Total number of users	249
Total number of posts	14198
Average ± Standard Deviation of the post of a user	57±98,7
Average ± Standard Deviation of the words on a post	26±54,56
Total number of words	369180
Size of the dictionary	32454
Users labelled as CONTROL	124
Users labelled as DEPRESSION	125

Table 3.8: Test set instance details.

In 3.8b and 3.7b box-plots, there are some users who have a ridiculous amount of posts compared to the other users. We have decided to limit their number of post with an upper bound. The bound will be an approximation of the third quartile of the bounding box. In the case of the ***Test*** set **200** posts and in the case of the train set **450** posts (taking the first post until bound). Difference of characteristics of the data-set can be seen after limiting the

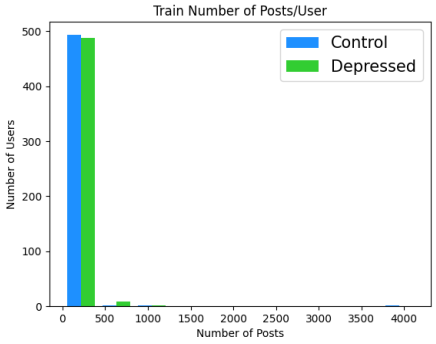


(a) Histogram: number of users (Y) who wrote a given number of posts (X) in the Test set.

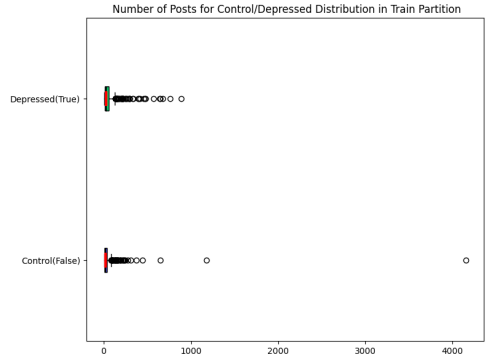


(b) Box plot of number of posts per user in the ***Test*** set.

Figure 3.7: Plots relevant for understanding the ***Test*** set.

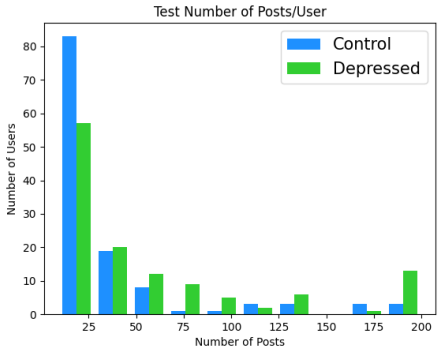


(a) Histogram: number of users (Y) who wrote a given number of posts (X) in the Train set.

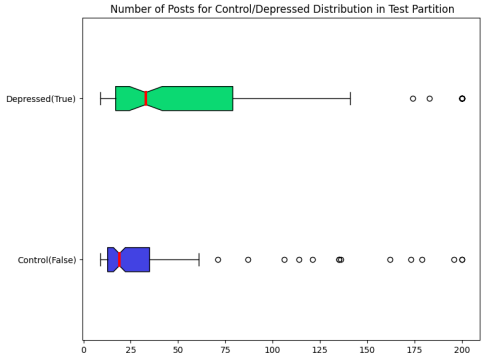


(b) Box plot of number of posts per user in the ***Train*** set.

Figure 3.8: Plots relevant for understanding the ***Train*** set.

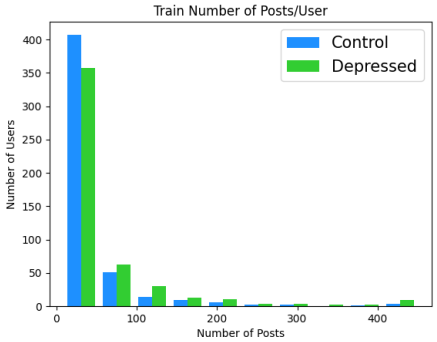


(a) Histogram: number of users (Y) who wrote a given number of posts (X) in the ***Test*** set after bounding post numbers.

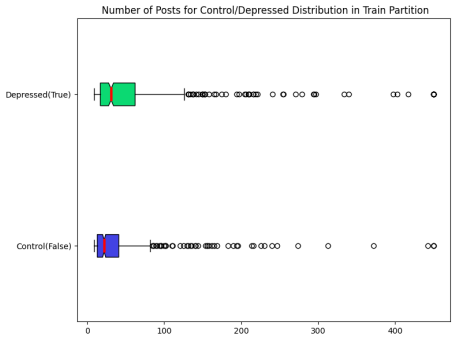


(b) Box plot of posts per user in the ***Test*** set after bounding posts.

Figure 3.9: Plots relevant for understanding the ***Test*** set after bounding post numbers.



(a) Histogram: number of users (Y) who wrote a given number of posts (X) in the ***Train*** set after bounding post numbers.



(b) Box plot of number of posts per user in the ***Train*** set after bounding post numbers.

Figure 3.10: Plots relevant for understanding the ***Train*** set after bounding post numbers.

Training Set	
Total number of users	993
Total number of posts	49854
Average ± Standard Deviation of the post of a user	50,2±70,89
Average ± Standard Deviation of the words on a post	26,16±52,64
Total number of words	1304466
Size of the dictionary	67255
Users labelled as CONTROL	497
Users labelled as DEPRESSION	496

Table 3.9: ***Training*** set instance details after bounding post numbers.

***Test* Set**	
Total number of users	249
Total number of posts	11735
Average ± Standard Deviation of the post of a user	47,12±52,7
Average ± Standard Deviation of the words on a post	27,23±57,52
Total number of words	319641
Size of the dictionary	29823
Users labelled as CONTROL	124
Users labelled as DEPRESSION	125

Table 3.10: ***Test*** set instance details after bounding post numbers.

number of posts in tables 3.9 and 3.10, they can be observed also in the figures 3.9b and 3.10b.

3.1.2.2 Annotations

The process of annotation is quite thorough, it goes as the following: The sequences of more than five SuicideWatch for a single user were divided into multiple annotation units containing up to five posts each. For example, a user with 12 posts would yield three annotation units of their first 5 posts, next 5 posts, final 2 posts. In order to determine user-level risk, the experts considered a user to have the highest risk associated with any of their annotation units.

The experts defined a four-way categorization of risk adapting [20] work (who provided lay definitions based on risk categories of the Thomas E. Joiner work [21]):

- **No Risk** (or "None"): They do not see evidence that this person is at risk for suicide.
- **Low Risk**: There may be some factors here that could suggest risk, but They do not really think this person is at much of a risk of suicide.
- **Moderate Risk**: They see indications that there could be a genuine risk of this person making a suicide attempt.
- **Severe Risk**: They believe this person is at high risk of attempting suicide in the near future.

The experts then defined two sets of annotator instructions. The ***short*** instructions, intended only for experts, simply presented the above categorization and asked them to follow their training in assessing patients with suicide risk. A ***long*** set of instructions was similar in intent to emulate the work mentioned in [20], but whereas their instructions focused on three risk factors (thoughts of suicide, planning, and preparation), they identified four families of risk factors:

1. *thoughts* includes not only explicit ideation but also, e.g., feeling they are a burden to others or having a "fuck it" (screw it, game over, farewell) thought pattern.
2. *feelings* includes, e.g., a lack of hope for things to get better, or a sense of agitation or impulsivity (mixed depressive state).
3. *logistics* includes, e.g., talking about methods of attempting suicide (even if not planning), or having access to lethal means like firearms.
4. *context* includes, e.g. previous attempts, a significant life change, or isolation from friends and family.

In both sets of instructions, annotators were also asked to label the post (if there are more than one) that most strongly supports the judgment, and they were told that choices should never be downgraded.

3.1.2.3 Expert Annotation

The organizers selected 245 users at random to create a set of 250 annotation units that were labelled independently by four volunteer experts in assessment of suicide risk.

1. Suicide prevention coordinator for the Veteran's Administration
2. Co-chair of the National Suicide Prevention Lifelines Standards, Training and Practices SubCommittee
3. Doctoral student with expert training in suicide assessment and treatment whose research is focused on suicidality among minority youth
4. Clinician in the Department of Emergency Psychiatry at Boston Childrens Hospital

Two of these experts received the detailed ***long*** instructions, and the other two were given the ***short*** instructions.

3.1.2.4 Crowdsourced Annotation

The organizers created a task on CrowdFlower (crowdflower.com) using the long instructions. They also restricted participation to high performance annotators (as determined by the CrowdFlower platform) and who also agreed with our annotations on seven clear test examples. Although the organizers began with 1,097 users to annotate, Crowdsourcer participation tailed off at 934. After discarding any annotation unit labelled by fewer than three annotators, our data comprises 865 users and 905 annotation units. THey used CrowdFlower's built-in consensus label as the crowdsourced label for each unit. In both cases, they generated a user-level consensus label using the Dawid-Skene (1979) model for discovering true item states/effects from multiple noisy measurements (Passoneau and Carpenter, 2014; see discussion in [18] work).

They recommend using the crowdsourcing train/test split for direct comparison with 2019 shared task papers, and using the full expert-annotated data-set for final testing, since the expert annotations have strong inter-rater reliability.

3.2 Methodology

In this section, we will specify a working methodology to maximize efficiency during the thesis development and enhance productivity. The working methodology and the way to work must follow these steps:

- Research and establish a solid knowledge-base about the background of depression detection and its relationship with suicide ideation.
- Search and identify the appropriate corpus (one or more) for the task.
- Analyse the appropriate corpus (one or more) for the task.
- Develop and document the code, sequentially, with the guidance of the advisors. Given the nature of the tasks at hand, it is expected that we may need to make changes in the development process to analyse different outcomes. Therefore, an iterative work methodology will be employed.
- Analyse the results and draw conclusions.

- Ask any doubts or questions to the advisors to continuously stay on track and keep up the workflow.

By following this methodology, we aim to streamline the workflow, optimize productivity, and ensure effective management of the thesis tasks.

3.2.1 Procedure

The methods we have chosen are common in the Natural Language Processing field, many researches involve this processes in addition to Machine Learning techniques, such as [3] [8]. Natural Language Processing (NLP) involves the use of computational techniques to analyse, understand, and generate human language. The steps involved (see figure 3.11) in NLP (see [22]) can vary depending on the specific task and approach, the ones we have used are:

- Data Acquisition
- Text Preprocessing
- Model Selection and Feature Extraction
- Model Training and Evaluation

The code we have used is inspired in the starting code of CLPysch2021, it can be found in GitHub [23].

3.2.1.1 Data Acquisition

The first step is to gather the required data for the NLP task. The data should be representative of the problem domain and large enough for training and evaluation. As said, we have collected text from existing data-sets composed by social media posts.

3.2.1.2 Text Preprocessing

Once the data is acquired, it needs to be preprocessed to prepare it for further analysis. The preprocessing steps we have applied include:

- Removing irrelevant information like HTML tags, special characters, UTF-8 emoji codes...
- Word Tokenization: The most common form of tokenization is word tokenization, which involves splitting a text into individual words. For example, given the sentence "Just posted a photo" the word tokenization would produce the tokens: ["Just", "posted", "a", "photo"].
- Converting text to lowercase for standardization.
- Removing stop words (common words like "the" or "and").

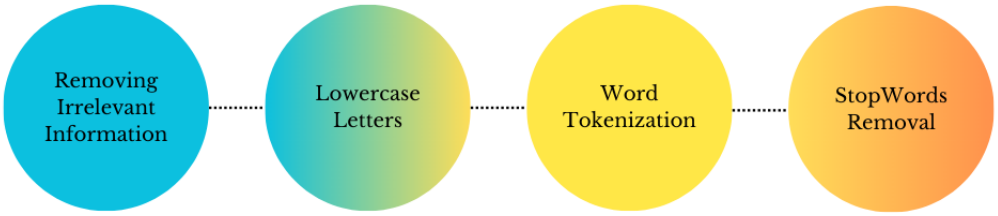


Figure 3.11: Visual Steps of text preprocessing.

3.2.1.3 Model Selection and Feature Extraction

Over the vast amount of possibilities, most researchers are prone to use Logistic Regression in our type of research [3] [11][9] [10]. Naturally, it is the one we have used. Logistic Regression (LR) is a linear classification approach used to estimate the probability occurrence of binary response based on one or more predictors and features. It is explained in [24], originally in article [25].

We will extract the features from our processed text (posts of each user) by converting them into a word n-gram with the help of `CountVectorizer` library from `sklearn.feature_extraction.text`. An n-gram refers to a contiguous sequence of n words from a given text, the "n" in "n-gram" represents the number of words in the sequence. There are various researchers that use this technique to extract features [11] [3].

Now, given training instances described as feature vectors $x_i \in \mathbb{R}^n$, ($i \in \{1, \ldots, l\}$ with l being the size of the training sample) in two classes (i.e. $c \in \mathcal{C} = \{1, -1\}$), and a vector $\mathbf{y} \in \mathbb{R}^l$ such that $y_i = \{1, -1\}$, represents the real class-value for the instance x_i, a linear classifier is modelled as weight vector $\mathbf{w}$. The decision function made by the classifier is the sign function, thus, the estimated class, y_i is obtained as in expression (3.1).

$$y_i(x_i) = sign(w^T x_i) \tag{3.1}$$

It is named "logistic" because it uses a logistic or sigmoid function to model the probability of an input belonging to a particular class as expressed in (3.2).

$$log(1 + e^{-y_i w^T x_i}) \text{ or } \frac{1}{1 + e^{-y_i \boldsymbol{w}^T x_i}} \tag{3.2}$$

- y_i: Class of the sample i.
- x_i: Vector of features (input) of the sample i.
- w: Weight vector that the model has learnt.

It is a simple though efficient model if apart from classifying we desire to acquire the probability of such classification. In other words, Logistic Regression returns the probability of our observation (x) being positive [26]. The probability model of logistic regression is shown in equation (3.3).

$$P(y \mid \boldsymbol{x}) = \frac{1}{1 + e^{-y \boldsymbol{w}^T \boldsymbol{x}}}, \text{ where } y = \pm 1, \tag{3.3}$$

So probabilities for two-class classification are immediately available. For a k-class problem, we need to couple k probabilities, a more advance heuristic is needed, explained in [26].

Luckily, our case does not involve such heuristic, thus we only have two labels to classify, if the user has depression or not. Thus, we will use the expression (3.3) to calculate the probability.

In our case, if the model returns a user with probability ≥ 50 we assume that he/she has depression. In the contrary, if the model returns a user with probability < 50 we assume that he/she does not have depression.

3.2.1.4 Model Training and Evaluation

We have particularly used the LIBLINEAR solver, mentioned in [26], to train the model. This particular solver adds L2 regularisation to the equation (see equation 3.4), increasing the generalisation and correcting overfitting.

$$\min_{\boldsymbol{w}} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{l}\log\left(1+e^{-y_i\boldsymbol{w}^T\boldsymbol{x}_i}\right) \tag{3.4}$$

The model is then trained using labelled data, where the input features and corresponding labels, in our case depression and control labels, are used to learn the underlying patterns in the data. The LIBLINEAR solver (the one we employ), uses line-search Newton method to optimize the model outcome. To have a main idea of the type of algorithm we are talking about, we will explain the base of the line-search Newton method, Trust Region Newton Method. The line-search Newton method is too advanced and complicated to fit in this work.

At each iteration of a Trust Region Newton method for minimizing $f(\mathbf{w})$, we have an iterate $\mathbf{w}^k$, a size Δ_k of the trust region, and a quadratic model, see equation 3.5.

$$q_k(\mathbf{s}) = \nabla f\left(\mathbf{w}^k\right)^T\mathbf{s} + \frac{1}{2}\mathbf{s}^T\nabla^2 f\left(\mathbf{w}^k\right)\mathbf{s} \tag{3.5}$$

As the approximation of the value $f\left(\mathbf{w}^k+\mathbf{s}\right) - f\left(\mathbf{w}^k\right)$. Next, we find a step $\mathbf{s}^k$ to approximately minimize $q_k(\mathbf{s})$ subject to the constraint $\|\mathbf{s}\| \leq \Delta_k$. We then update $\mathbf{w}^k$ and Δ_k by checking the ratio of the actual reduction in the function to the predicted reduction (see equation 3.6) in the quadratic model.

$$\rho_k = \frac{f\left(\mathbf{w}^k+\mathbf{s}^k\right) - f\left(\mathbf{w}^k\right)}{q_k\left(\mathbf{s}^k\right)} \tag{3.6}$$

The direction is accepted if ρ_k is large enough where $\eta_0 > 0$ is a pre-specified value, see equation 3.7.

$$\mathbf{w}^{k+1} = \begin{cases} \mathbf{w}^k + \mathbf{s}^k & \text{if } \rho_k > \eta_0, \\ \mathbf{w}^k & \text{if } \rho_k \leq \eta_0, \end{cases} \tag{3.7}$$

From Lin and Moré (1999) [27], updating rules for Δ_k depend on positive constants η_1 and η_2 such that $\eta_1 < \eta_2 < 1$, while the rate at which Δ_k is updated relies on positive constants

σ_1, σ_2, and σ_3 such that $\sigma_1 < \sigma_2 < 1 < \sigma_3$. The trust region bound Δ_k is updated by the rules

$$\begin{array}{lll} \Delta_{k+1} \in \left[\sigma_1 \min\left\{\left\|\mathbf{s}^k\right\|, \Delta_k\right\}, \sigma_2 \Delta_k\right] & \text{if} & \rho_k \leq \eta_1, \\ \Delta_{k+1} \in \left[\sigma_1 \Delta_k, \sigma_3 \Delta_k\right] & \text{if} & \rho_k \in (\eta_1, \eta_2), \\ \Delta_{k+1} \in \left[\Delta_k, \sigma_3 \Delta_k\right] & \text{if} & \rho_k \geq \eta_2. \end{array} \tag{3.8}$$

Trust Region Newton Method

Given $\mathbf{w}^0$.
For $k = 0, 1, \ldots$ (outer iterations)
  if $\nabla f\left(\mathbf{w}^k\right) = \mathbf{0}$, stop.
  Find an approximate solution $\mathbf{s}^k$ of the trust region sub-problem

$$\min_{\mathbf{s}} q_k(\mathbf{s}) \quad \text{subject to } \|\mathbf{s}\| \leq \Delta_k. \tag{3.9}$$

  Compute ρ_k via 3.6.
  Update $\mathbf{w}^k$ to $\mathbf{w}^{k+1}$ according to 3.7.
  Obtain Δ_{k+1} according to 3.8.

Algorithm 3.1: A trust region algorithm for logistic regression

After training the model, we evaluated the results to assess its performance and generalization ability. For the results we have used various metrics that symbolize the efficiency of the model, explained afterwards. The model's performance is measured on a separate test set that was not used during training to ensure an unbiased evaluation.

3.2.2 Evaluation Metrics

In our work we have used a total of six different metrics, explained in [28] [29]:

- Confusion Matrix.
- AUC score.
- Precision.
- Recall.
- False Positive Rate.
- F-score.

Previous research, such as, [3][4][11] show that the model efficiency is represented correctly.

3.2.2.1 Confusion Matrix and Variations

A confusion matrix, see table 3.11, often referred to as a "confusion table," is a fundamental tool in the field of machine learning and statistics used to evaluate the performance of a classification model, particularly in the context of supervised learning. It provides a

clear summary of how well a classification model is performing by comparing the model's predictions to the actual ground truth. A confusion matrix typically consists of a grid with four key components:

- True Positives (TP): These are cases where the model correctly predicted the positive class. In other words, the model correctly identified instances that actually belong to the class being predicted.
- True Negatives (TN): These are cases where the model correctly predicted the negative class. It indicates that the model correctly identified instances that do not belong to the class being predicted.
- False Positives (FP): Also known as Type I errors, these are cases where the model incorrectly predicted the positive class when the actual class is negative. In other words, the model made a false alarm, incorrectly identifying instances as belonging to the positive class.
- False Negatives (FN): Also known as Type II errors, these are cases where the model incorrectly predicted the negative class when the actual class is positive. It means the model missed instances that actually belong to the positive class.

		Assignment	
		$+$	$-$
Label	$+$	TP	FN
	$-$	FP	TN

Table 3.11: Confusion matrix representation.

In our case, the positive class is the one that HAS depression, therefore the negative one is the one we identify as non depressive.

Confusion matrices are particularly valuable for understanding the performance of binary classification models. They serve as the basis for various performance metrics such as accuracy, precision, recall, F1-score, and the Receiver Operating Characteristic (ROC) curve. See table 3.12 for some of the simpler cases.

3.2.2.2 F-score

The F-score, also known as the F1-score, is a widely used performance metric in binary classification that combines the precision and recall of a model into a single value. It's particularly useful when there is an uneven class distribution (like in our first data-set), which can make accuracy a misleading metric. The F-score is calculated as the harmonic mean of precision and recall, as stated in (3.10).

$$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.10}$$

There are variations of the F-score, such as the F-beta score, which allows to place more emphasis on either precision (for $F_\beta > 1$) or recall (for $F_\beta < 1$) depending on the specific needs of each problem.[29]

Metrics	Formula	Evaluation Focus
Precision (p)	$\frac{tp}{tp+fp}$	Precision is used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class.
Recall (r)	$\frac{tp}{tp+fn}$	Recall is used to measure the fraction of positive patterns that are correctly classified
False Positive Rate (fpr)	$\frac{fp}{fp+tn}$	It is a measure of the model's tendency to incorrectly classify negative instances as positive. In other words, it quantifies the rate at which the model makes false positive predictions among all the actual negative instances.

Table 3.12: Information and formula of the metric that are involved in our work.

$$F_\beta = \left(1+\beta^2\right)\frac{pr}{r+\beta^2 p} = \frac{\left(1+\beta^2\right)TP}{\left(1+\beta^2\right)TP+\beta^2 FN+FP} \tag{3.11}$$

3.2.2.3 Area under the ROC Curve (AUC)

One common summary statistic derived from the ROC curve is the Area Under the Curve (AUC). The AUC quantifies the overall performance of the model; a higher AUC indicates better discrimination between positive and negative instances. An AUC of 0.5 suggests random guessing, while an AUC of 1.0 indicates perfect classification. It is calculated as the following:

$$AUC = \frac{S_p - n_p\left(n_n+1\right)/2}{n_p n_n}$$

where, S_p is the sum of the all positive examples ranked, while n_p and n_n denote the number of positive and negative examples respectively. The AUC was proven theoretically and empirically better than the accuracy metric [30] for evaluating the classifier performance and discriminating an optimal solution during the classification training.

3.2.3 Evaluation methods

In our first data-set, we had some issues regarding the data. First of all, the given test partition was too small to even consider it usable, therefore we decided to split the training data into *train* and *dev* partitions. But, another main issue surfaced: the training data classes were imbalanced, consequently we had to split the data accordingly. We did it by partitioning the data stratifically (used in [31]). In other words, it preserves the relative proportions of different classes or categories within the data when splitting it.

Here's how a stratified split works:

- Count the Classes: First, identify the different classes or categories in your data-set. In binary classification, typically there are two classes (positive and negative), but in multi-class classification, there may be more.

- Calculate Class Proportions: Determine the proportion or percentage of each class in the entire data-set. For example, if there is a binary classification problem, calculate the percentage of positive and negative examples in the data-set.
- Maintain Proportions: When splitting the data-set into subsets (e.g., training and testing sets), make sure that each subset contains the same relative proportion of each class as the original data-set. This ensures that both subsets are representative of the overall class distribution.

CHAPTER 4

Experimental Results

In this chapter, we embark on a comprehensive journey through the outcomes of meticulously designed experiments, each orchestrated to scrutinize a facet of our thesis framework. These experiments encompass diverse data-sets, controlled scenarios, and a range of evaluation metrics that collectively unearth the nuances of our study's outcomes. The insights gleaned from this empirical voyage illuminate not only the attainment of our work goals but also the insights that propel us toward deeper understanding. The experiments divide into two sections, one for each data-set. They both facilitate the process and journey of our methods applied to the indicated data-set.

4.1 CLPSych2021 Practice data-set results

We have divided the original training data into two disjoint subsets: **train** and **dev** in a stratified way (for further details on stratification, turn to 3.2.3, specifically, to page 36). The partition, composed by 20% of the original training data assigned to **dev** and 80% assigned to **train**. We used the **dev** partition to test the model and its efficiency after training it with the **train** partition (see page 19, sub-section 3.1.1.2).

As mentioned in the description of the data-set, section 3.1.1, the results are given by the model in boolean(true/false), but we use Control(False) and Depressed(True) as a more appealing way to understand the type of users to be predicted.

Metrics	
F1-score	0.711
True Positive Rate/Recall	0.80
False Positive Rate	0.562
Precision	0.64
AUC score	0.641

Table 4.1: Metrics of the results from ***Dev*** partition.

The results shown in the figure 4.1 indicate that almost 50%(0.44) of the labeled with depression correctly. It seems that, it is harder to correctly classify the users that do not

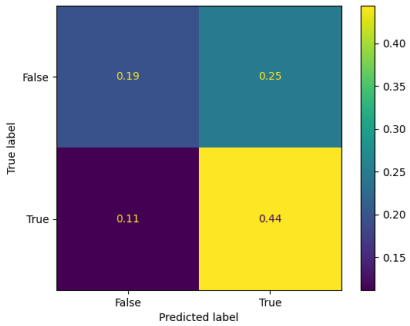


Figure 4.1: Confusion Matrix of ***Dev***.

have depression (Control). The metrics that make use of the Confusion Matrix can be seen in table 4.1.
We can see there is a vast amount (0.25) of Depressed predicted user, but that are actually Control users. Nevertheless, in the other way around, we have 0.11 Control predicted user that are actually Depressed users.

The results shown until now can be conceived as a base-line, as a mere binary classification. We made use of all the available data, all the tweets of each user. As more data is usually better for prediction, the results shown in table 4.1 represent the peak performance of the current model. To continue our work, to go further in our investigation, we thought about early depression detection. Some of the following questions arise:

- In practice, could we make an announcement before we have all the posts from a user?
- Would the quality of the prediction be greatly degraded by reducing the input information?

After some thought, we came up with an idea to explore early depression detection. In the data-set (see sub-section 3.1.1.2), each tweet comes with the date in which it was created. We thought that ordering the tweets chronologically and only using a determined amount iteratively would be an interesting path to follow.
To explore the ability of the model to assess the users with just a restricted amount of information and, thus, to assess early prediction. We wondered how would the model drop predictive ability providing a limited amount of tweets in their chronological order. To this end we designed two experimental scenarios. The following subsections gather the results of each scenario respectively. The metrics in the following subsections are limited to the F1-Score, Precision and AUC score. We chose them because they are the most essential to evaluate the model and too many metric figures are more confusing or unhelpful.

4.1.1 Variation A

We fixed a threshold similar to the mean of the number of tweets in each user, 25 tweets. We discarded the users that had less than that number of tweets. This lead to the reduction of the amount of users in each set by half more or less. With the users with more than 25 tweets we assessed the performance of the model. We started using the first 5 tweets of

each user as input information and continuing up to 25 tweets, increasing 5 tweets in each experiment, so 5 different measures were computed. As specified, in each iteration, we take into account five more tweets of each user, in a chronological order. That way, we hope to find a threshold/point where, with X number of tweets, we can state that the user has depression. In other words, to detect as early as possible if a user has depression or not. We applied this variation in two ways:

1. Only modifying the ***Dev*** partition. This is, the users from the ***Dev*** partition have limited amount of tweets.

2. Modifying the ***Dev*** and ***Train*** partitions simultaneously. This is, the users from the ***Dev*** and ***Train*** partition have limited amount of tweets.

We restrict the users to those with, at least, 25 tweets. To get a better understanding of the users that were left here is some information about them, check table 4.2.

Information of Variation A users		
	Dev	Train
Total number of users	16	61
Users labelled as CONTROL	7	28
Users labelled as DEPRESSION	9	33

Table 4.2: Information about the users in Variation A

In this variation, the amount of users in each iteration does not change. Each user type is still balanced and stratified. The progression of the number of tweets used in each iteration is a matter of multiplying the users by the number of tweets used in each iteration. In both ways of using this variation the tweets of the ***Dev*** partition are reduced (check the table 4.3 to see the progression).

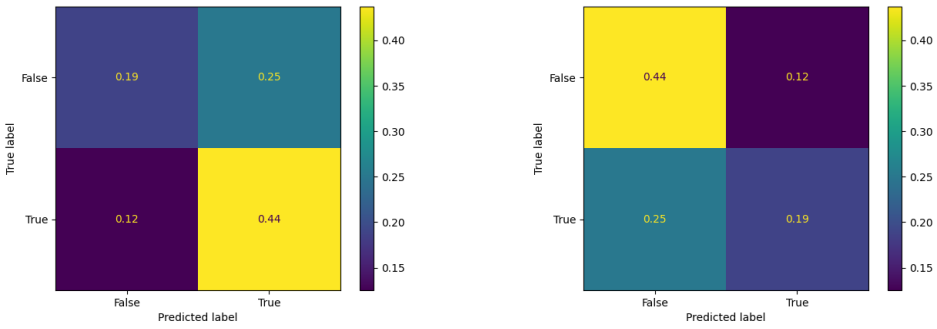
Information of Variation A iterations in *Dev*					
Iterations	1	2	3	4	5
Tweets used per user	5	10	15	20	25
Total tweets	80	160	240	320	400
Total words	734	1363	2019	2518	3101
Average $\pm$ Standard Deviation of words in one tweet	9$\pm$8	8.5$\pm$7	8.5$\pm$7	8$\pm$6.5	8$\pm$6
Dictionary length	543	953	1287	1571	1866

Table 4.3: Information about the changes of the ***Dev*** partition through the iterations of Variation A

We have measured the efficiency in both ways of using the variation, chronologically increasing five tweets per user in each iteration. The following sections show the results of the two modification types mentioned in variation A: modifying ***Dev*** and modifying ***Dev*** & ***Train*** at the same time.

4.1.1.1 Only modifying the *Dev* partition

The ***train*** set keeps all the tweets while the ***dev*** partition adds progressively the tweets in chronological order. The results observed in figure 4.2 show that depression class results are much better classified. It can be observed in the progression of the metrics, see figure 4.3, that the scores increase linearly at the first 4 iterations overall. Then, in the 5^{th} iteration the scores reach like what it seems a limit and don't change or even worsen like in the case of AUC score in figure 4.3b.



(a) Confusion matrix of the Depression class.

(b) Confusion matrix of the Control class.

Figure 4.2: Confusion matrix of the 5^{th} iteration only modifying the ***Dev*** partition.

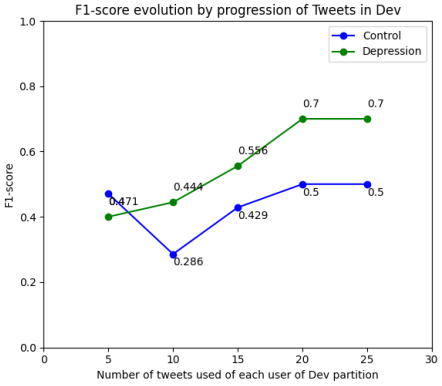
4.1.1.2 Modifying the *Dev* and *Train* partitions simultaneously

In this way of using variation A, the ***Train*** partition tweets are also reduced, a quick overview of the information per iteration is in the table 4.4.

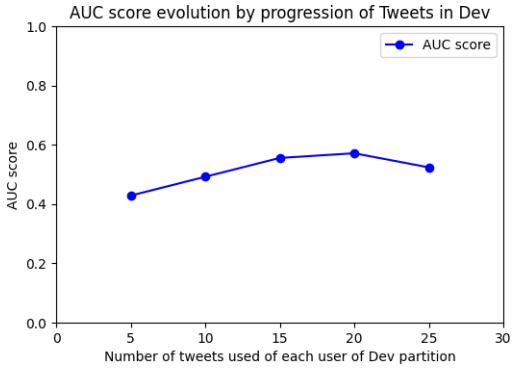
Information of Variation A iterations in *Train*					
Iterations	1	2	3	4	5
Tweets used per user	5	10	15	20	25
Total tweets	305	610	915	1220	1525
Total words	2377	4874	7424	9904	12415
Average $\pm$ Standard Deviation of words in one tweet	8$\pm$5.5	8$\pm$5.5	8$\pm$5.5	8$\pm$6	8$\pm$6
Dictionary length	1564	2702	3733	4521	5318

Table 4.4: Information about the changes of the ***Train*** partition through the iterations of Variation A. The number of users in each iteration does not change in Variation A, see table 4.2.

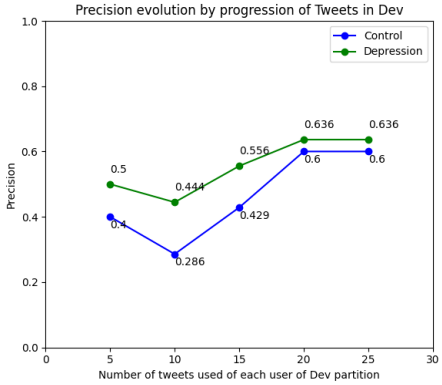
This time, the results of the model are worse than the previous ones. It is logical, as the training data we use is also smaller. The results observed in figure 4.4 show that depression class results are closer to the Control class. Although the results in general are worse, it can be noticed in figure 4.5 that there is a peak of the scores in the 4^{th} iteration where we use 20 tweets per user.



(a) F1-Score progression through iterations.

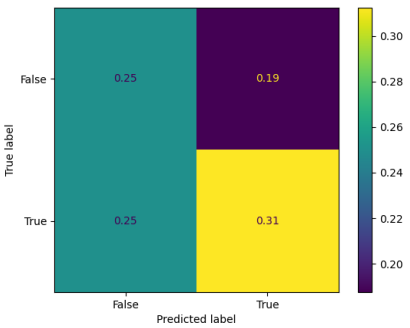


(b) AUC score progression through iterations (same in both classes).

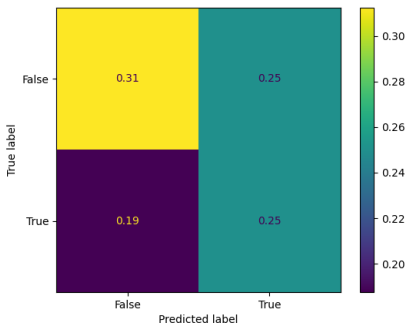


(c) Precision score progression through iterations.

Figure 4.3: Progression of metrics, only modifying the ***Dev*** partition in variation A.

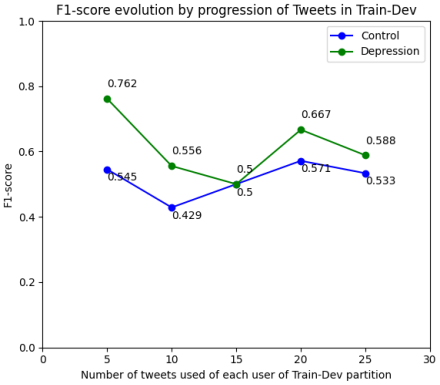


(a) Confusion matrix of the Depression class.

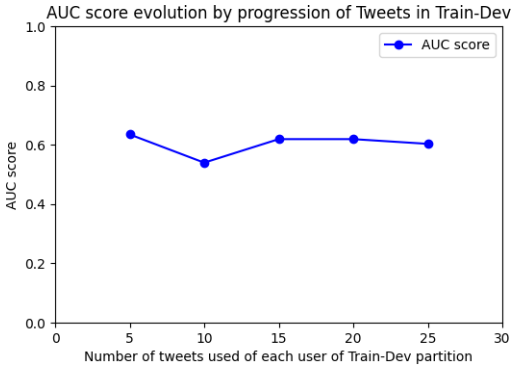


(b) Confusion matrix of the Control class.

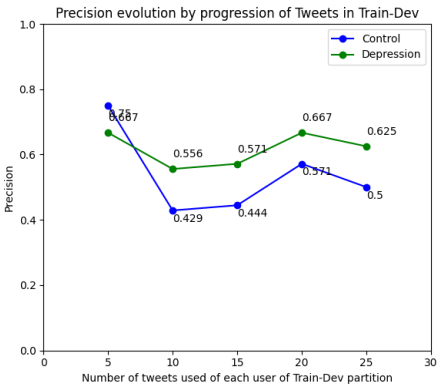
Figure 4.4: Confusion matrix of the 5^{th} iteration modifying the ***Dev*** and ***Train*** partitions simultaneously.



(a) F1-Score progression through iterations.



(b) AUC score progression through iterations (same in both classes).



(c) Precision score progression through iterations.

Figure 4.5: Progression of metrics, only changing ***Dev*** and ***Train*** partitions in variation A.

4.1.2 Variation B

We fixed a threshold similar to the mean of the number of tweets in each user, 25 tweets. With the users that were left we started to measure the efficiency of the model, but there was a problem, not all users had 25 tweets. So we thought that the users that did not have the amount of tweets required would use all the tweets they had, instead of discarding them. We started using the first 5 tweets of each user (if the user had fewer tweets the max available) and continuing until 25 tweets increasing 5 tweets each time, so 5 different measures/iterations were computed. So we also applied this variation in two ways:

1. Only modifying the ***Dev*** partition. This is, the users from the ***Dev*** partition have limited amount of tweets.

2. Modifying the ***Dev*** and ***Train*** partitions simultaneously. This is, the users from the ***Dev*** and ***Train*** partition have limited amount of tweets.

In this variation, the amount of users involved in testing the model does not change.

In the cases where the number of tweets to be used exceeds the number of tweets that the user has, we will use all the tweets of the user. To overview the changes of the data in ***Dev*** partition in each iteration, look at table 4.5.

Information of Variation B iterations in *Dev*					
Iterations	1	2	3	4	5
Total tweets	171	326	449	556	646
Total words	1417	2568	3513	4300	4866
Average ± Standard Deviation of tweets in one user	5±1	9±2	12.5±4	15.5±6	18±8
Average ± Standard Deviation of words in one tweet	8±6.5	8±6	8±6	8±6	7.5±6
Dictionary length	1054	1751	2178	2494	2771

Table 4.5: Information about the changes of the ***Dev*** partition through the iterations of Variation B.

4.1.2.1 Only modifying the *Dev* partition

Overall, the results are a bit smoother as we have more tweets from the start of the iteration. This has sense because none of the users are discarded, so almost all the tweets are used. As in variation A, the 5^{th} iteration the scores reach like what it seems a limit and don't move or even worsen like in the case of AUC score in figure 4.6b.

4.1.2.2 Modifying the *Dev* and *Train* partitions simultaneously

In this way of using variation B, the ***Train*** partition tweets are also reduced, a quick overview of the information per iteration is in the table 4.6.

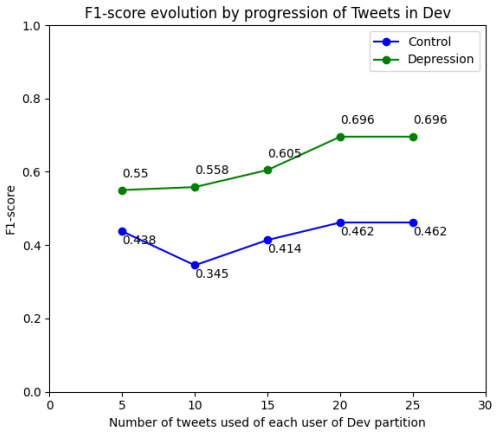
Information of Variation B iterations in *Train*					
Iterations	1	2	3	4	5
Total tweets	676	1295	1833	2254	2581
Total words	5439	10458	15073	18708	21385
Average ± Standard Deviation of tweets in one user	5±1	9±2	13±4	16±6	18±7.5
Average ± Standard Deviation of words in one tweet	8±6	8±6	8±6.5	8±6.5	8±6
Dictionary length	3222	5209	6715	7722	8453

Table 4.6: Information about the changes of the ***Train*** partition through the iterations of Variation B.

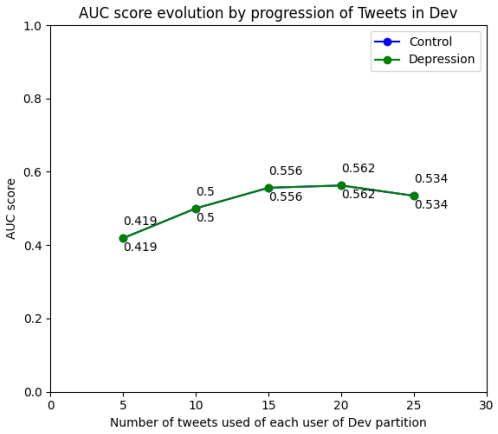
As for the results (see figure 4.7), they are a bit worse than in the general case, smoothed by the increase of tweets. In the beginning there is an unusual behaviour in the scores, but it straightens as the iterations go by.

4.2 CLPSych2021 Reddit data-set results

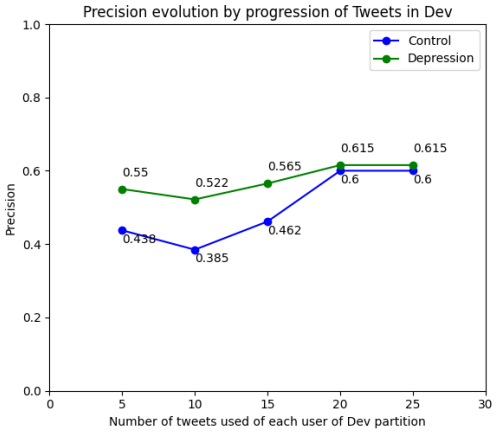
The following results were obtained by processing the Version 2 of the Reddit Suicidality Data-set 3.1.2. As the data is much larger, the results are more generalised than the previous data-set. We should mention that as the data is large there are some problems with the variance (number of posts per user) but we solved them previously as explained in 3.1.2.1 of



(a) F1-Score progression through iterations.



(b) AUC score progression through iterations (same in both classes).



(c) Precision score progression through iterations.

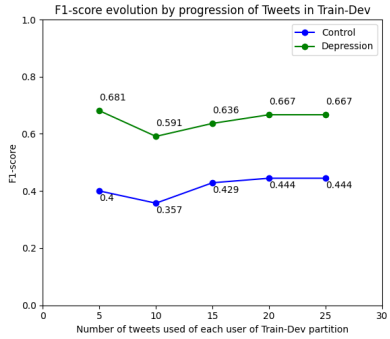
Figure 4.6: Progression of metrics, only modifying the ***Dev*** partition in variation B.

the *Materials* section. We used the ***Test*** set to test the model and its efficiency after training it with the ***Train*** set. The main results are shown in 4.7.

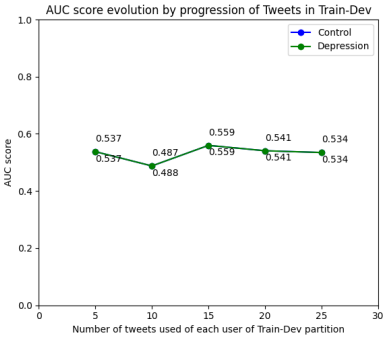
The method of evaluation is the same as the previous data-set, with variations A/B applying each one of them to the ***Train***/***Test*** partitions. Although, due to the size of the actual data-set, there are some changes in the length and threshold of the methods.

4.2.1 Variation A

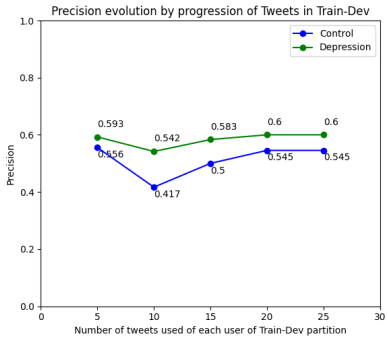
This time we fixed the threshold similar to the mean of the number of posts in each user, 55 posts. We discarded the users that had less than that proportion number of posts, that reduced the amount of users in each set by half more or less. With the users that were left, we started to measure the efficiency of the model. We started using the first 5 posts of each user and continuing until 55 posts, increasing 5 posts each time, so 11 different



(a) F1-Score progression through iterations.



(b) AUC score progression through iterations (same in both classes).



(c) Precision score progression through iterations.

Figure 4.7: Progression of metrics, only changing ***Dev*** and ***Train*** partitions in variation B.

Metrics	
F1-score	0.816
True Positive Rate/Recall	0.744
False Positive Rate	0.081
Precision	0.903
AUC score	0.847

Table 4.7: Metrics of the results from CLPsych2021 Reddit crowd data-set.

measures/iterations were computed. As specified, in each iteration, we take into account five more posts of each user, in a chronological order. So we also applied this variation in two ways:

1. Only modifying the ***Test*** set. This is, the users from the ***Test*** partition have limited amount of posts.

2. Modifying the ***Train*** and ***Test*** sets simultaneously. This is, the users from the ***Test*** and ***Train*** partition have limited amount of posts.

To get a better understanding of the users that check the information about them in table 4.8. In this variation, the amount of users in each iteration does not change. Each

Information of Variation A users		
	Test	Train
Total number of users	61	224
Users labelled as CONTROL	17	88
Users labelled as DEPRESSION	44	136

Table 4.8: Information about the users in Variation A in the CLPSych2021 Reddit Data-set.

user class was balanced before restricting the posts, but after the restriction there are more Depression users than Control users. That means that overall the Depressed users have more posts (average±StD of posts per users is 62,6±99,1) than Control users (average±StD of posts per users is 50,2±199,1). The standard deviation of Control users is much higher than Depression users. Thankfully, it is fixed with the reduction of the variation we have applied, as it flattens the difference of posts per user.
The class imbalance affects the results significantly, even some of the outcome is not expected and has little sense. This setback is handled in the following sections 4.2.1.1 and 4.2.1.2, individually. Although, when only modifying the ***Test*** set there
shouldn't be any unusual results, as the ***Train*** set is perfectly balanced and as Gary King says [32]:

> *"In my experiments I have found that sometimes there can be a bias in favour of the minority class, but that is caused by wild over-fitting."*

The progression of the number of posts used in each iteration is a matter of multiplying the users by the number of posts used in each iteration. In both ways of using this variation the posts of the ***Test*** partition are reduced, check the table 4.9 to see the progression.

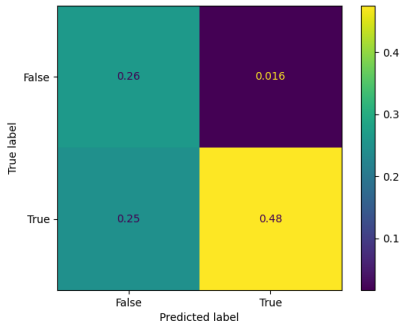
Information of Variation A iterations in *Test* (1/2)					
Iterations	1	2	3	4	5
Posts used per user	5	10	15	20	25
Total posts	305	610	915	1220	1525
Total words	7615	14439	22969	29330	36907
Average ± StD of words	24.9±76.1	23.7±64.2	25.1±67.7	24±63.1	24.2±61.9
Dictionary length	2984	4621	5961	7020	8196

Information of Variation A iterations in *Test* (2/2)						
Iterations	6	7	8	9	10	11
Posts used per user	30	35	40	45	50	55
Total posts	1830	2135	2440	2745	3050	3355
Total words	45048	52537	60756	70345	81457	90293
Average ± StD of words	24.6±67.3	24.6±59.3	24.9±58.8	25.6±60.1	26.7±63.6	26.9±63.6
Dictionary length	9221	10191	11096	12133	13296	14061

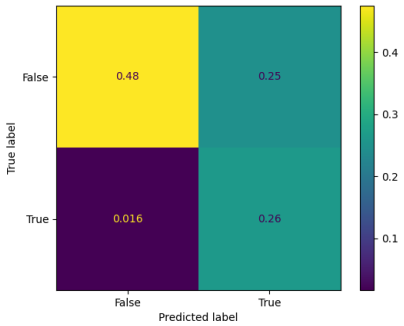
Table 4.9: Information about the changes of the ***Test*** partition from the CLPSych2021 Reddit Data-set through the iterations of Variation A.

4.2.1.1 Only modifying the *Test* set

The ***Train*** set keeps all the posts, while the ***Test*** set adds progressively the posts in chronological order. The class imbalance mentioned earlier affects the results in a way that they don't seem comprehensive and lack of sense. Various figures show that if we regulate the number of users, regarding their class (Depression or Control), the results are more coherent. To be more specific, we reduced the number of Depression class users to the first 17. The results with the user distribution 17(C)-44(D) (as in table 4.8) are shown in figure 4.10 and with the modified distribution 17(C)-17(D) the results are shown in figure 4.11. The difference can also be observed in the confusion matrices 4.8 and 4.9. Overall, the as the change on the results does not affect the final efficiency of the model.

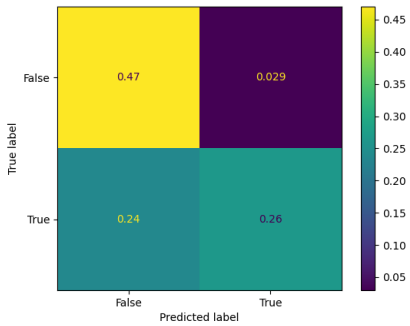


(a) Confusion matrix of the Depression class.

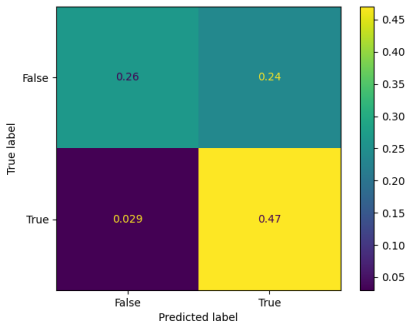


(b) Confusion matrix of the Control class.

Figure 4.8: Confusion matrices of the 11^{th} iteration only modifying the ***Test*** set in CLPSych2021 Reddit Data-set.



(a) Confusion matrix of the Depression class.

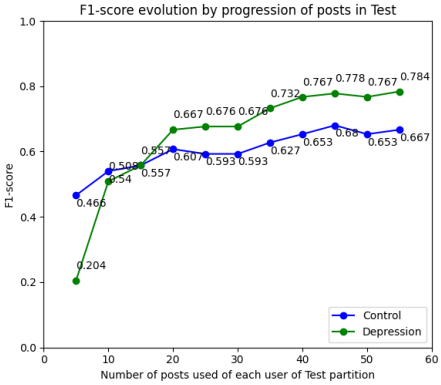


(b) Confusion matrix of the Control class.

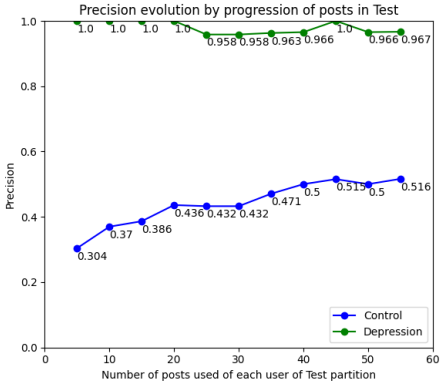
Figure 4.9: Confusion matrices of the 11^{th} iteration only modifying the ***Test*** set in CLPSych2021 Reddit Data-set after balancing the number of users of both classes to 17.

4.2.1.2 Modifying the *Train* and *Test* set simultaneously

In this way of using variation A, the ***Train*** set posts are also reduced, a quick overview of the information per iteration is in the table 4.10. This time the data imbalance applies to the ***Train*** and ***Test*** sets, as it is shown in table 4.8. The initial results are a bit abnormal as

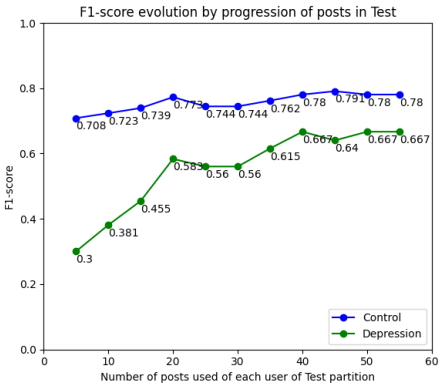


(a) F1-Score progression through iterations.

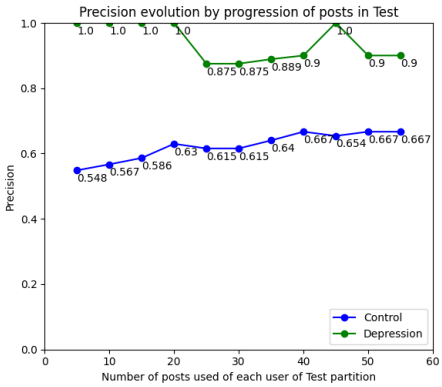


(b) Precision score progression through iterations.

Figure 4.10: Progression of metrics only modifying the ***Test*** set in variation A in CLPSych2021 Reddit Data-set.



(a) F1-Score progression through iterations.



(b) Precision score progression through iterations.

Figure 4.11: Progression of metrics only modifying the ***Test*** set in variation A in CLPSych2021 Reddit Data-set after balancing the number of users of both classes to 17.

Depression class metrics are much higher than the Control class results, see figure 4.13. Consequently, we applied different solutions to get more realistic results:

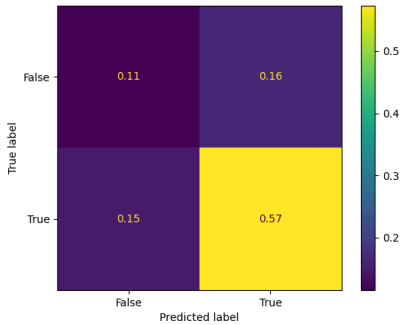
- Figure 4.14: Leveling the users of each class in ***Train*** set: 88 users of each class.
- Figure 4.15: Leveling the users of each class in ***Train*** set and ***Test*** set: 88 user for ***Train*** and 17 users for ***Test*** of each class.

The second solution gives more realistic results, as the two classes have similar results, with a slight peak at the 4^{th} iteration.

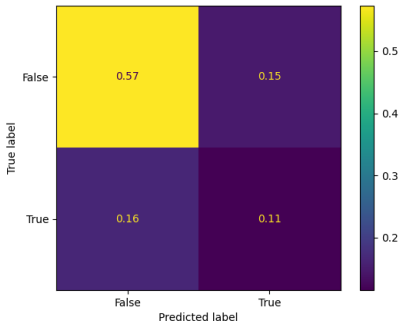
Information of Variation A iterations in *Test* (1/2)					
Iterations	1	2	3	4	5
Posts used per user	5	10	15	20	25
Total posts	1120	2240	3360	4480	5600
Total words	24503	48189	70022	96623	121539
Average ± StD of words	21.8±46.7	21.5±44.2	20.8±40.8	21.5±42.6	21.7±42.7
Dictionary length	7143	10763	13430	16213	18478

Information of Variation A iterations in *Test* (2/2)						
Iterations	6	7	8	9	10	11
Posts used per user	30	35	40	45	50	55
Total posts	6720	7840	8960	10080	11100	12220
Total words	146474	170766	197119	225792	249275	275570
Average ± StD of words	21.8±42.2	21.8±41.5	22±41.7	22.4±42.5	22.2±41.8	22.3±42.1
Dictionary length	20327	22042	23736	25445	26778	28317

Table 4.10: Information about the changes of the ***Train*** partition from the CLPSych2021 Reddit Data-set through the iterations of Variation A.

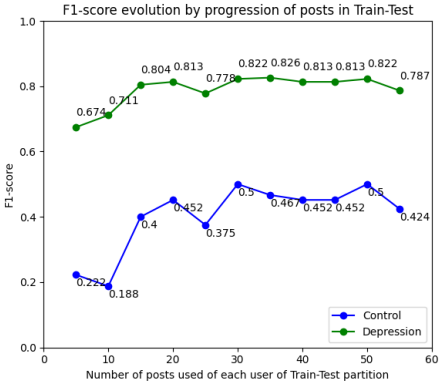


(a) Confusion matrix of the Depression class.

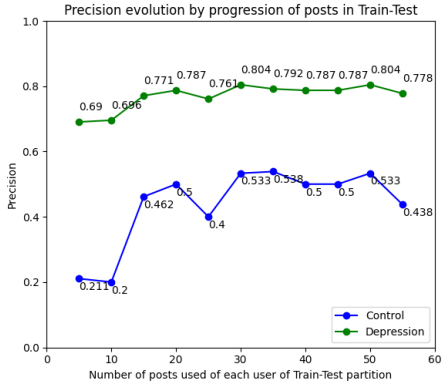


(b) Confusion matrix of the Control class.

Figure 4.12: Confusion matrices of the 11^{th} iteration only modifying the ***Train*** and ***Test*** set in CLPSych2021 Reddit Data-set.

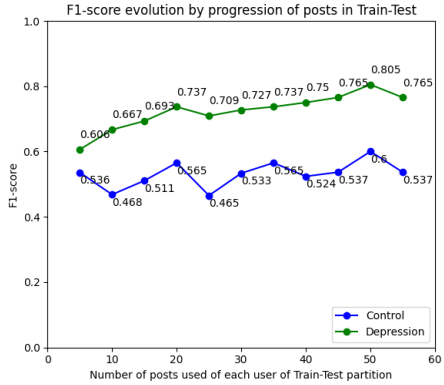


(a) F1-Score progression through iterations.

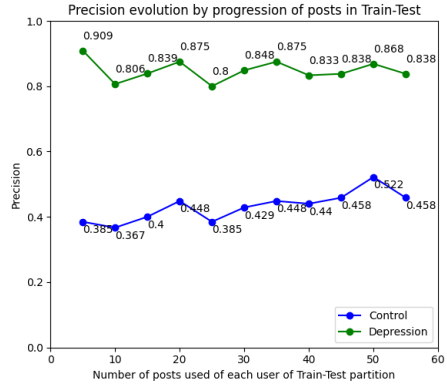


(b) Precision score progression through iterations.

Figure 4.13: Progression of metrics only modifying the ***Train*** and ***Test*** set in variation A in CLPSych2021 Reddit Data-set.



(a) F1-Score progression through iterations.

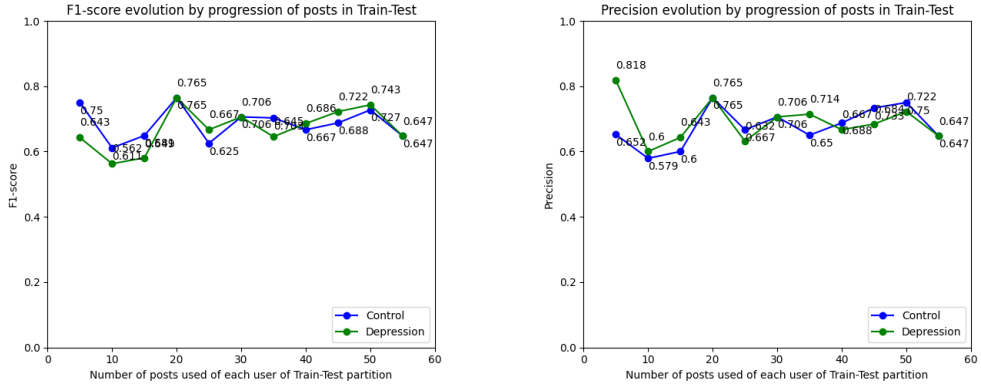


(b) Precision score progression through iterations.

Figure 4.14: Progression of metrics only modifying the ***Train*** and ***Test*** set in variation A in CLPSych2021 Reddit Data-set after balancing the number of users of both classes to 88.

4.2.2 Variation B

In this variation the amount of users involved in testing the model does not change, derivatively there is no class imbalance involved here. We use the posts of all the users, even the ones that do not have 55 posts. In those cases, when the number of posts to be used exceeds the number of posts that the user has, we will use all the posts of the user instead of discarding them.
We start using the first 5 posts of each user (if the user had fewer posts the max available) and continuing until 55 posts increasing 5 posts each time, so 11 different measures/iterations were computed.
So we also applied this variation in two ways:



(a) F1-Score progression through iterations.

(b) Precision score progression through iterations.

Figure 4.15: Progression of metrics only modifying the ***Train*** and ***Test*** set in variation A in CLPSych2021 Reddit Data-set after balancing the number of users of both classes to 17 for ***Test*** set and 88 for ***Train*** set.

1. Only modifying the ***Test*** set. This is, the users from the ***Test*** partition have limited amount of posts.

2. Modifying the ***Train*** and ***Test*** sets simultaneously. This is, the users from the ***Test*** and ***Train*** partition have limited amount of posts.

In the two cases the ***Test*** set is modified, so to overview the changes of the data in the Test set in each iteration, look at table 4.11.

Information of Variation B iterations in *Test* (1/2)					
Iterations	1	2	3	4	5
Total posts	1245	2481	3518	4313	4941
Total words	35256	71788	101702	123640	140733
Average ± StD of posts	5±0	9.9±0.2	14.1±1.7	17.3±3.7	19.8±5.7
Average ± StD of words	28.3±76.1	28.9±63.5	28.9±62.8	28.6±61.1	28.4±60.6
Dictionary length	8679	13150	15706	17446	18741

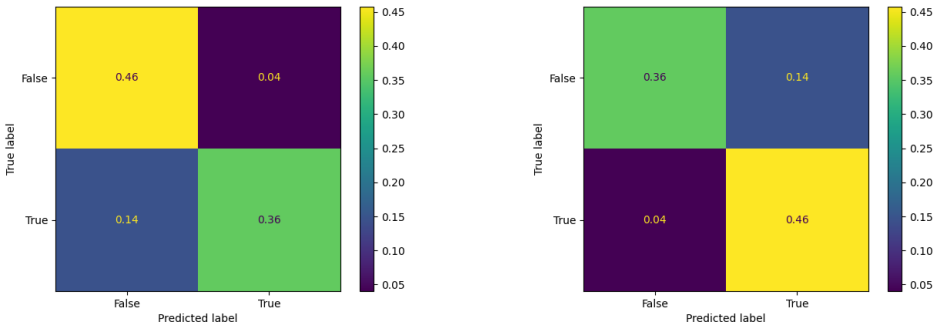
Information of Variation B iterations in *Test* (2/2)						
Iterations	6	7	8	9	10	11
Total posts	5493	5978	6402	6785	7133	7455
Total words	153830	166715	176986	187628	199736	209020
Average ± StD of posts	22±7.8	24±9.9	25.7±11.9	27.2±13.8	28.6±15.7	29.9±17.5
Average ± StD of words	28±59	27.9±58.5	27.6±57.9	27.6±58.2	28±59.7	28±59.9
Dictionary length	19679	20617	21339	22086	22892	23413

Table 4.11: Information about the changes of the ***Test*** partition from the CLPSych2021 Reddit Data-set through the iterations of Variation B.

4.2.2.1 Only modifying the *Test* set

The ***train*** set keeps all the posts, while the test adds progressively the posts in chronological order. It can be observed in the progression of the metrics (see figure 4.17) that the scores are higher for depression class regarding precision but slightly lower regarding F1-score.

Overall, the results are higher than the previous, see figure 4.18. In the progressive results, we noticed that the iterative process converge/stabilize, indicating that there is minimal variation or change observed as the iterations progress.



(a) Confusion matrix of the Depression class.

(b) Confusion matrix of the Control class.

Figure 4.16: Confusion matrices of the 11^{th} iteration only modifying the ***Test*** set in CLPSych2021 Reddit Data-set.

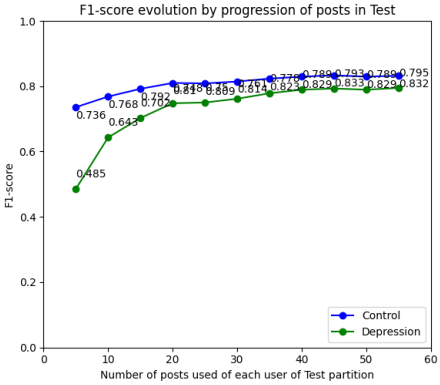
iteratively

4.2.2.2 Modifying the *Train* and *Test* set simultaneously

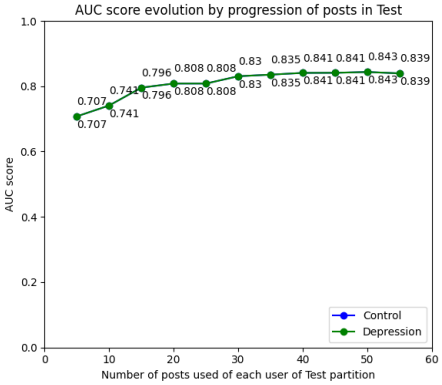
In this way of using variation B, the ***Train*** set posts are also reduced, a quick overview of the information per iteration is in the table 4.12. To look at the results, check figure 4.19. Once again the Depression class takes the lead in precision, it is logical as the model has more data of the Depression class to learn from. This time, the progression of the results changes a bit more during the iterations. It can be observed that the Control class overdoes the Depression class in figure 4.18.

4.3 Data examples

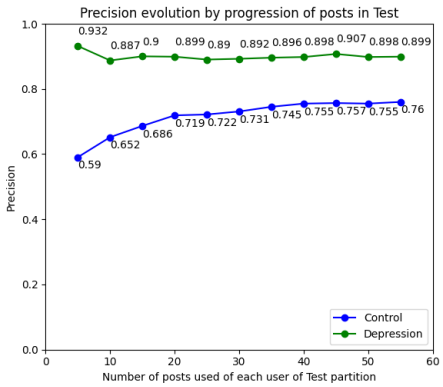
After creating the baseline model from the data, we output a results file. The results file is a Tabular Separated Value (TSV) file with the following form:



(a) F1-Score progression through iterations.

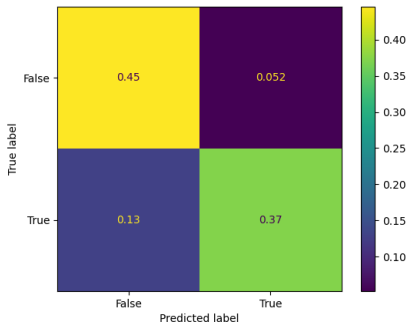


(b) AUC score progression through iterations (same in both classes).

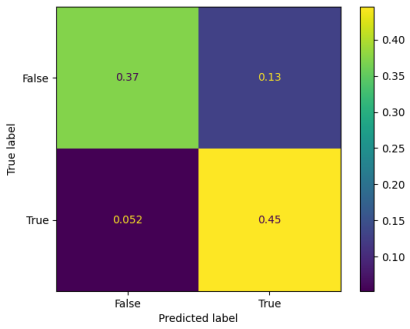


(c) Precision score progression through iterations.

Figure 4.17: Progression of metrics only modifying the ***Test*** set in variation B in CLPSych2021 Reddit Data-set.



(a) Confusion matrix of the Depression class.



(b) Confusion matrix of the Control class.

Figure 4.18: Confusion matrices of the 11^{th} iteration modifying the ***Train***/***Test*** set simultaneously in CLPSych2021 Reddit Data-set.

Information of Variation B iterations in *Train* (1/2)					
Iterations	1	2	3	4	5
Total posts	4965	9868	13931	17289	20100
Total words	141318	276703	378200	471880	547095
Average ± StD of posts	5±0	9.9±0.2	14±1.9	17.4±3.9	20.2±6
Average ± StD of words	28.5±61.2	28±59.5	27.1±57.2	27.3±57	27.2±56
Dictionary length	19310	27924	33415	37845	41241

Information of Variation B iterations in *Train* (2/2)						
Iterations	6	7	8	9	10	11
Total posts	22528	24603	26358	27880	29227	30406
Total words	610062	661625	706151	747352	778960	807423
Average ± StD of posts	22.7±8	24.7±10	26.5±11.9	28±13.7	29.4±15.5	30.6±17.1
Average ± StD of words	27±55.6	26.9±55	26.8±54.6	26.8±54.6	26.6±54.1	26.5±53.9
Dictionary length	43817	45998	47729	49161	50248	51332

Table 4.12: Information about the changes of the ***Train*** partition from the CLPSych2021 Reddit Data-set through the iterations of Variation B.

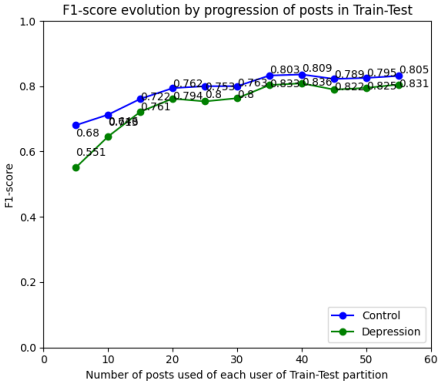
```
[USER_ID] \t [LABEL] \t [SCORE]
```

Where `USER_ID` is the ID field from the source file, `LABEL` is either Depressed or Control, and `SCORE` is a real-valued score output from our system, where larger numbers indicate the Depressed class and lower numbers indicate Control.

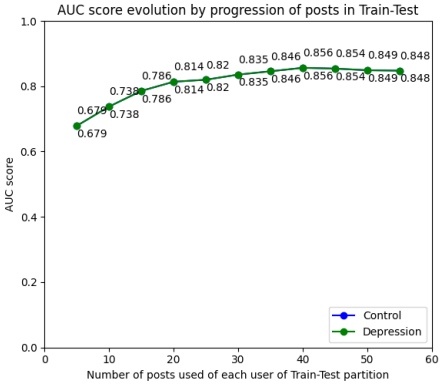
To show it more clearly, here are some examples of two users of each end of the prediction and their texts.

- User: *3426279078* - Prediction: *True 0.8098802908067035*
  This user has the following tweets:
  - {"id": "1213057857379135488",
    "created_at": "2020-01-03 11:21:45 UTC",
    "text": "let's spit on those words and force us down.
    follow my IG: @delgacomarkanthony like/follow my FBpage:Anthony Delgaco"}
  - {"id": "1219625533559828480",
    "created_at": "2020-01-21 14:19:21 UTC",
    "text": "drama mo sis,e ang dami mo naman kachat haha"}
  - {"id": "1215073061881303041",
    "created_at": "2020-01-09 00:49:27 UTC",
    "text": "it's just something I've learned to be kind
    even when they do it wrong. photography:||\ud83c\udfa5: sir
    John Agustin location: BGC,taguig city   follow my IG: @delgacomarkanthony
    follow/like my fbPage: Anthony delgaco"}

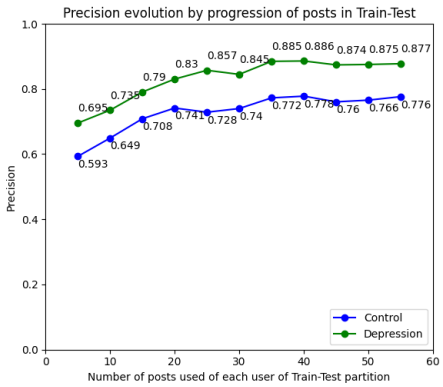
- User: *860781311111442434* - Prediction: *False 0.4433833181235666*
  This user has the following tweets:
  - {"id": "1314201175781072904",



(a) F1-Score progression through iterations.



(b) AUC score progression through iterations (same in both classes).



(c) Precision score progression through iterations.

Figure 4.19: Progression of metrics modifying the ***Train***/***Test*** set simultaneously in variation B in CLPSych2021 Reddit Data-set.


```
  "created_at": "2020-10-08 13:49:11 UTC",
  "text": "'\u062e\u062f\u0627 \u0645\u0648\u062c\u0648\u062f \u06c1\u06d2
  \u062a\u0648 \u0628\u062c\u0627\u060c \u0627\u06af\u0631 \u0646\u06c1\u06cc\u06ba
  \u062a\u0648 \u0627\u0633\u06d2 \u062a\u062e\u0644\u06cc\u0642
  \u06a9\u0631\u0646\u0627 \u0686\u0627\u06c1\u06cc\u06d2\"
  ~\u0648\u0627\u0644\u0679\u06cc\u0626\u0631  2-year-old Zainab
  Raped Murdered Body mutilated (Chest &amp; Abdomen Slit)
  \ud83d\ude2d \ud83d\udc94  \"  https://t.co/7rWSxQDNvp"}

- {"id": "1239233937538506753",
  "created_at": "2020-03-15 16:56:09 UTC",
  "text": " All Exams Cancelled  All students will be promoted to next class by
  Duckworth Lewis Method\ud83d\ude02\ud83e\udd23\ud83d\ude1c"}

- {"id": "1273915133597712389",
  "created_at": "2020-06-19 09:46:50 UTC",
  "text": "Now Modi &amp; Amit Shah r preparing Veer Chakra award for these Soldiers
  Like Abinandhan Wing Commander of IAF. This is a big big blot on Modi Govt. &amp;
  Godi Media.Shame on u"}
```

As mentioned, it seems that the prediction of depression of the users is not quite accurate in comparison to their tweets. It is true that the first user uses words such as:

- Spit
- Force us down
- Drama
- Wrong

But the sentences as a whole do not have anything to do with depression.
The second user has words like:

- Award
- Shame
- Murdered
- Mutilated

Surprisingly, such words like *murder* or *mutilated* do not have any effect on the prediction, as it stands less than 0.5. In this case, the user does not have depression, but it should have been predicted with higher chance of having depression due to the used words.

4.4 Error Analysis

In the practice data-set, in both Variations the Depression class scores are superiors to the Control class ones, these might be due to a little class imbalance in the ***Dev*** partition. On the whole, the results from Variation B are higher, it is rational as there are more tweets in the training process. In order to gain a more generalised model that will have a smaller true risk; true risk and empirical risk are well illustrated in the book [24].

In the CLPSych Reddit data-set, the amount of input data, compared with the previous data-set, was larger and in Machine Learning it makes a big (positive) difference. Furthermore, if the input data that we possess is balanced, we have the advantage to work with raw and pure data from each class equally. That is rarely the case, in an actual real case scenario minority classes appear all the time. In those cases, new data is created in order to level the minority classes with others (data augmentation, re-sampling...) or class weight are used to notify the model about the imbalance (the model punishes higher/more weight the errors in the minority class). We mentioned the difference of the input data because this data-set had far more posts or quantity than the previous one, see tables 3.7 and 3.8. As previously stated, the outcome has a direct relation with the input, thus the results of this data-set are better overall (see table 4.7). Because of the stated fact related to data quantity, we can say that the model behaves in the way we expected, this is, correctly and logically.

In addition, in Variation A we reduce the users to the ones that have $\geq$55 posts (see table 4.8). The results that came out after modifying only the ***Test*** set are not very logical, the figure 4.10a shows that the Depression class is fully correctly classified almost every time. It is not a realistic approach as we would have achieved the perfect model. In addition, the control class had terrible results.
In order to fix it, we tried levelling the ***Test*** set each class users to 17 (see figure 4.11). This time the results are a bit more normal, but not comprehensive at all.
Furthermore, when modifying the ***Train*** and ***Test*** sets simultaneously we also have some data imbalance issues, but we managed to fix them (see figure 4.15). We noticed a slight peak at results of the 5^{th} while using 20 posts per user, the two classes reach to 0.76 of precision and F1-score. It can be considered an achievement, only with 20 posts per user we can say with 0.76 of confidence that the user has depression. We state that 20 posts can be a threshold for identifying individuals with depression with 0.76 of confidence (see figure 4.15).

4.5 Discussion

During the whole work we managed to apply a variety of methods on the acquired data. The results we have gained after the experiments are diverse, engaging and sometimes a bit perplexes. That in mind, we will mention the results drawn from each data-set we used, Reddit Data-set and Practice Data-set.

4.5.1 CLPSych2021 Practice data-set

We have reached the conclusion that quantity of the data in this data-set is too small to think of the results as a real case scenario, this is, generalized.
In table 4.1 shows the False Positive rate or fallout (explained in table 3.12) is very high. It represents the proportion or percentage of negative instances that are mistakenly identified as positive. As it mentions in [33]:

> "*We find motivation for the use of evaluation metrics that assess bad recommendations, along with (or complementarily to) metrics that assess good. Simple metrics involving false positives, such as fallout [34] ... can suitably meet this purpose; they are defined as:*"

With that being said, we can assure that when the model has to predict a Control user it will do it incorrectly with a probability of 0.56. As mentioned, it is too high to be acceptable.

Because of the poor data, the overall precision of the model is mediocre, although it is admirable to reach 0.64 of accuracy with this little data.

Regarding the different variations of using the data, we can determine by looking at the figures that in both Variations, when modifying the two partitions, the initial F1-score and precision (see figures 4.5/4.7) are very high in both classes. This can mean that with very little data or even in the range of 5-10 tweets we can get a score that is even more accurate than the score with all the tweets. Conversely, we might be wrong, this effect is due to the little data and is a matter of casualty.

Finally, due to the low data that we manage in this data-set we cannot find any relation or assumption about the efficiency of the model and the chronological addition of tweets, this is, early detection of depression.

4.5.2 CLPSych2021 Reddit data-set

We will start this subsection by admitting that in Machine Learning, it makes a big difference in the outcome the amount of input data we obtain. Furthermore, if the input data that we possess is balanced, we have the advantage to work with raw and pure data from each class equally. That is rarely the case, in an actual real case scenario minority classes appear all the time. In those cases, new data is created in order to level the minority classes with others (data augmentation, re-sampling, transferred learning...) or class weight are used to notify the model about the imbalance (the model punishes higher/more weight the errors in the minority class).

We mentioned the difference of the input data because this data-set had far more posts or quantity than the previous one, see tables 3.7 and 3.8. As previously stated, the outcome has a direct relation with the input, thus the results of this data-set are better overall (check table 4.7).
The fallout is less than 0.01 which is a big difference in comparison with the previous data-set. Precision is exactly 0.9 which is astonishing and recall is fairly high 0.74.

Variation B results are quite standard, and we have not seen any interesting results coming from them.

Finally, we want to remark that using only the first 20 posts (chronologically speaking) per user in both ***Train*** and ***Test*** sets, the precision and F1-score are 0.76 (see figure 4.15) which compared to the initial (and best) results it is quite impressive. We can conclude that with 0.764 probability over 1 the model will predict correctly only using the first 20 chronological posts of the users.

4.5.3 Comparison with Related Work

We will compare the related work with the Reddit Data-set results as there are the best among the two data-sets.

The precision (in table 4.7), we obtain with the Reddit data-set is of 0.903, it is considerably high for the model we are using (LR). It is even better than the average precision results obtain from [35] research. It is even more impressive because the mentioned work uses more advanced models and techniques, such as word embeddings (word2vec), Long Short-Term Memory (LTSM) + Recurrent Neural Network (RNN), Convolutional Neural Network (CNN)...We must admit that even if the precision is higher the F1-Score is lower that the mentioned research.

Regarding AUC score the [11] states the following:

> "*Our best Multi-Task Learning (MTL) model predicts potential suicide attempt, as well as the presence of atypical mental health, with AUC > 0.8.*"

In comparison, our AUC score (see table 4.7) is 0.847. A bit more higher that the mentioned study and again, the mentioned study has used a bit more advanced. In the

study they also use a Logistic Regression model as in our case and their AUC score when predicting depression users is 0.763. Considerably lower than ours.

CHAPTER 5

Conclusions

To conclude the work, we have drawn several conclusions regarding the work accomplished and the tools utilized. Furthermore, we have gleaned valuable lessons that we deem significant for our future professional and academic endeavours. In addition, we have identified areas for improvement and outlined potential future avenues of exploration.

5.1 Conclusions

We will proceed to develop the conclusions following the main goals that we established at the beginning of the thesis.

5.1.1 First Main Goal: Framework Development

This main goal entailed implementing a NLP framework. For that we had to achieve sub-goals.

First, we had to make some background research in order to gain some knowledge about the field. Many research and events exist about detecting depression, anxiety and similar mental problems, thus it was straightforward. More information can be seen in chapter 2, in the *Background* section.

Following up on the work we searched data/data-sets/corpus and analyzed it. We focused on social media data sources for the research and found many data-sets that could be used for our work. From all the options, we started working on the practice data-set and then extended the analysis to a larger and more diverse data-set. Overall, the data management was really successful. The data-sets we have used can be found in the 3. chapter, in the *Materials* section. The data we have not used can be found in the *Appendix*.

The depression estimation has been done correctly. We defined a probability threshold for identifying individuals with depression. Then, we fed social media posts to the model for training, that way the model learned how to. Before feeding the data to the model we used NLP techniques to pre-process the text. We have also used NLP techniques to train and obtain results. Further information can be found in the *Methodology* section on the 3. chapter.

Finally, we analyze the results and tried to understand them. It can be seen in the 4. chapter, in the last two sections (*Error Analysis* and *Discussion*).

Mostly, we conclude that we successfully achieved to complete the 100% of the main task, designing and implementing a NLP framework.

5.1.2 Second Main Goal: Early Detection

This main goal is related with developing early detection techniques.

During this challenge our approach consisted of using data in chronological order. In other words, we monitored changes in results as new data is added sequentially to the model. That is, adding data in the training process and/or testing process. In summary, we experimented with various methods and variations of posts to diversify results. We tried to define a posts threshold for identifying individuals with depression with some confidence, although the evidence is not quite reliable. The experiments can be found at the beginning of the 4. chapter, where the results are presented by means of plots and tables.

Overall, we consider this main goal to be completely done. In spite of the uncertainty on defining a posts threshold for identifying individuals with depression.

5.1.3 Third Main Goal: Project management

We are pleased to report that all of the sub-objectives we set out to achieve have been successfully realized throughout the course of this thesis.

Firstly, we ensured the proper management of the thesis, meticulously allocating resources, maintaining thorough documentation, and, most importantly, reaching a robust and well-supported conclusion.

Secondly, ongoing refinement of thesis objectives were instrumental in providing clarity and guidance to our thesis efforts. We had our main goals in mind, but sometimes we had to ensure we were on track and aligned with our goals by reviewing our next steps. Moreover, by embracing an iterative work methodology, we demonstrated our ability to adapt to changing thesis needs and outcomes.

Thirdly, we can proudly present a comprehensive thesis document that encapsulates our methodology, findings, and conclusions.

Fourthly, we diligently developed and documented our thesis code in a sequential manner, incorporating valuable input and insights from our advisors.

Lastly, our commitment to maintaining open and transparent communication with our advisors played a pivotal role in our success. Their guidance and clarification throughout the thesis were indispensable, ensuring we stayed on the right path and made informed decisions.

In summary, we almost achieved all of these sub-objectives and they have been instrumental in the successful completion of our thesis.

5.2 Project management

This section is centered in showing the development of the work regarding the initial tasks. That is, taking into account the hours that we estimated for each task and the real time it took to accomplish each task. To observe the difference in each task check table 5.1. The total estimated time for the thesis is only 20 hours less than the real one. It is quite a

		Estimated(h)	Actual(h)	Diff.
1	Thesis	424	444	+20
1.1	Management	58	55	-3
1.1.1	Brainstorming.	8	6	-2
1.1.2	Planning.	10	8	-2
1.1.3	Methods and Communication.	25	21	-4
1.1.4	Tracking progress.	15	20	+5
1.2	Research	45	43	-2
1.2.1	Search and read related works.	25	22	-3
1.2.2	Gather information.	10	10	0
1.2.3	Analyze strengths and weaknesses.	5	4	-1
1.2.4	Gain inspiration for work.	5	7	+2
1.3	Corpus	47	46	-1
1.3.1	Obtain corpus of social media posts.	10	11	+1
1.3.2	Analyze corpus.	12	15	+3
1.3.3	Get another corpus and analyze it.	25	20	-5
1.4	Development	143	148	+5
1.4.1	Select methodology.	8	5	-3
1.4.2	Apply methods to data.	15	12	-3
1.4.3	Early detection experiment.	35	40	+5
1.4.4	Result analysis.	15	11	-4
1.4.5	Code development.	70	80	+10
1.5	Conclusions	18	18	0
1.5.1	Gather the work outcome.	8	4	-4
1.5.2	Review the objective/goal accomplishment.	3	5	+2
1.5.3	Improvements of the work and acquired knowledge.	3	5	+2
1.5.4	Analyze the overall management of the work.	4	4	0
1.6	Documentation	113	134	+21
1.6.1	Write the Introduction chapter.	10	15	+5
1.6.2	Write the Related Work chapter.	15	12	-3
1.6.3	Write the Material and Methods chapter.	15	22	+7
1.6.4	Write the Experimental Results chapter.	25	30	+5
1.6.5	Write the Conclusions chapter.	15	10	-5
1.6.6	Write the Appendix chapter.	5	5	0
1.6.7	Reference other works.	8	5	-3
1.6.8	Correct the document.	10	25	+15
1.6.9	Make the presentation.	10	10	0

Table 5.1: Table that compares the hour estimation with the reality per task.

close gap, meaning most of the task hour estimation was more or less right. With some exceptions of course, like the Documentation work area and the development work area.

Regarding the Documentation work area, writing and explaining concepts was more difficult than anticipated. Apart from that, the some structure issues (tables, figures, floats...) were encountered because of the LaTeX syntax. The specific syntax made the writing process more tedious and slow. Other than that, the corrections also took more time than estimated, the attention to detail was remarkable.

Regarding the Development work area, the early depression detection section took more than it should. We dug deep in the chronological variations that we came up with(Variation A/B) and experimented with many different modifications, see chapter 4. Thinking and developing those variations took time. In addition, we had to code each variations in order to obtain the results, that took more time too.

Lastly, regarding the timeline (see figure 1.1), we were following quite precisely the estimated dates for each task. Nevertheless, due to work overload, we were not able to finish the work in June 2023, as it was predicted. We have finished the work in early September 2023.

5.3 Acquired knowledge

The following text recounts the valuable knowledge we gained from immersing in this work, which delved into the fascinating realms of machine learning techniques, Natural Language Processing (NLP), data-set search and analysis, data visualization, clear representation of results through plots and effective work management. This transformative journey enabled us to broaden our horizons and expand our understanding of these essential fields. Through the work's insights and experiences, we acquired a deeper appreciation for the practical applications and theoretical foundations that underpin these disciplines.

The thesis offered practical guidance on navigating the vast landscape of data-sets available on the internet. To achieve that, we learnt to search for them and proceeded with the necessary bureaucratic processes to lay a hand on them. It elucidated strategies for identifying relevant and reliable data-sets for research or analysis purposes. The thesis highlighted the significance of understanding data sources, metadata, and licensing agreements when selecting data-sets. In relation with that, we learnt about the importance of maintaining data integrity, ensuring compliance with privacy regulations, and acknowledging appropriate citations to promote transparency and reproducibility.

The exploration of visual data in the thesis demonstrated the power and impact of presenting data in a visually appealing manner. It emphasized the significance of data visualization as a means to effectively communicate complex concepts and patterns. By employing appropriate techniques, such as graphs, box plots, pie charts and histograms, one can enhance data comprehension and enable informed decision-making. The thesis highlighted the importance of selecting appropriate visual representations tailored to specific contexts, ensuring clarity and maximizing the impact of the information conveyed.

A central theme in the thesis was the art of clearly representing results. It provided valuable insights into the various types of representation, such as confusion matrices, tables and line graphs. The thesis emphasized the significance of choosing the most suitable plot type to present data, considering the nature of the variables and the relationships being

analysed. It underscored the importance of labelling axes and utilizing colours effectively to convey meaningful insights through plots, for example the results of each class shown in different colours.

The work also delved into the realm of machine learning, offering a comprehensive understanding of its principles and methodologies. It explored key concepts of supervised learning centred in Natural Language Processing (NLP). The work emphasized the significance of data preprocessing, model selection, and model evaluation to ensure optimal performance of machine learning algorithms. Additionally, it highlighted the ethical considerations surrounding bias, correctness, and transparency in machine learning systems. More specifically, the thesis provided valuable insights into the domain of Natural Language Processing (NLP). It shed light on the challenges associated with processing and understanding human language, including tasks such as depression detection and text classification. The thesis delved into techniques such as tokenization or stop-words removal, which enable effective NLP applications. It emphasized the potential of NLP in various fields, for example early detection of depression.

The thesis imparted invaluable wisdom on effective work management strategies within these domains. It stressed the importance of maintaining a structured approach, setting realistic goals, and adhering to thesis timelines. We learnt the significance of continuous learning, experimentation, and documentation to foster professional growth and knowledge sharing.

Through the thesis's rich narrative and the thesis's thought-provoking experiences, we gained a wealth of knowledge spanning clear visual of data, representing results through plots, machine learning, NLP, data-set search through the internet, and work management. This immersive journey not only expanded our theoretical understanding but also provided practical capabilities.

5.4 Enhancements and future improvements

We find it difficult to select the best ways to enhance the model/work as there are so many pathways to follow.
We would say that one of the best possibilities is to experiment with the posts of the users we take when reducing them to a certain amount. Instead of using the first posts of the users, chronologically, one alternative could be to use the ones starting from last, in some way inverse chronological order. It could be beneficial, as if the user has depression it probably will go escalating over time and the signs of depression will be greater. Another possibility is to use random posts of the users.

Some other way of experimenting can be related with the metrics. The metrics we have chose are Precision Recall, Fallout, F1-Score and AUC, instead of using those there are metrics that can be more helpful. Some of them are, Mean Average Precision (MAP), Perplexity or f-Latency.

Furthermore, we would suggest on trying new ways to squeeze the simultaneous modification of training and test set of Variation A, in section 4.2.1.2. In our opinion, there is an intriguing prospect that the method promises.

To finish, we will propose some possible intriguing ideas to improve the work:

- NLP representations or embeddings more accurate and reliable depression detection can be implemented [22], such as:
  - bag-of-words, where each word in the text is represented as a feature.
  - word embeddings (e.g., Word2Vec or GloVe), which capture semantic relationships between words, widely explained in [36].
  - Named Entity Recognition (NER), involves identifying and classifying named entities in text. A survey [37] covers approaches for tokenization, feature extraction, and classification techniques used for NER.
- Exploration of advanced feature engineering methods to capture linguistic nuances.
- Discussion of incorporating context-aware and temporal analysis for better understanding of dynamic emotional states.
- Using transfer learning approach for text classification [38].
- Upgrading the learning model, instead of a simple model like Logistic Regression, more advanced models can be used. This paper [39] proposes a model that combines Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNNs) for text classification.

Appendix

Useful Links

- The 2021 Computational Linguistics and Clinical Psychology initial workshop dataset description [18].
- The base code given in the 2021 Computational Linguistics and Clinical Psychology workshop, in GitHub [23].
- The code given in the 2021 Computational Linguistics and Clinical Psychology workshop to retrieve the practice dataset, in GitHub [14].
- An article that explains how to deal with data imbalance called *How to Improve Class Imbalance using Class Weights in Machine Learning* [40]. It clears up how to easily fix the problem with imbalanced data, with code included.
- Another project in GitHub [13] regarding depression detection, by Susan Wang, Labiba Kanij Rupty, Mahfuza Humayra Mohona, Aarthi Alagammai, Munira Omar and Marwa Qabeel.
- A LaTeX symbol and structure guide in PDF format called *The Comprehensive LATEX Symbol List* by Scott Pakin. [41]
- Mathpix: A web-based tool and software application that utilizes optical character recognition (OCR) and machine learning algorithms to recognize and convert mathematical equations and symbols from images into digital formats. It provides a convenient and efficient way for users to extract mathematical content from various sources such as textbooks, handwritten notes, or even screenshots. In summary, it can decompose a compiled LaTeX code (equations, formulas...) into the literal LaTeX code; even images!

Related National Challenges/Competitions

- CLPsych 2022: CLPsych has brought together researchers in computational linguistics and NLP, who use computational methods to better understand human language, infer meaning and intention, and predict individuals' characteristics and potential behavior, with mental health practitioners and researchers, who are focused on psychopathology and neurological health and engage directly with the needs of providers and their patients. This workshop's distinctly interdisciplinary nature

has improved the exchange of knowledge, fostered collaboration, and increased the visibility of mental health as a problem domain in NLP. The 2022 shared task introduced the problem of assessing changes in a person's mood over time on the basis of their linguistic content. For the purpose of the task they focused on posting activity in online social media platforms. In particular, given a user's posts over a certain period in time, they aim: (1) at capturing those sub-periods during which a user's mood deviates from their baseline mood – a post-level sequential classification task. They then build on this task, by leveraging it to further help them assess: (2) the risk level the user is at – a user-level classification task [1] & a continuation of the 2019 Shared Task [3]. Thus, the task consisted of the two subtasks: (1) the main task of identifying mood changes in users' posts over time and (2) the auxiliary task of showing how (1) helps them assess the risk level of a user. More about this event [42].

- eRisk 2019: eRisk explores the evaluation methodology, effectiveness metrics and practical applications (particularly those related to health and safety) of early risk detection on the Internet. The third task consisted on measuring the severity of the signs of depression, assessing the severity of the risk for depression. More about this event [43].

- eRisk 2018: eRisk explores the evaluation methodology, effectiveness metrics and practical applications (particularly those related to health and safety) of early risk detection on the Internet. The first task was Early Detection of Signs of Depression, that consisted on early predicting the signs of depression, as its name says.[44]

- BRAT: Example of how natural language processing techniques work in order to extract the needed information to perform a prediction. More about this event [45].

- CLPsych 2018: Also known as the Fifth Workshop on Computational Linguistics and Clinical Psychology. The main task of the year was Predicting Current and Future Psychological Health from Childhood Essays, as social media data-sets were still difficult to get. More about this event [46].

- CLPsych 2017: Also known as the Fourth Workshop on Computational Linguistics and Clinical Psychology. Same as the other workshops but with limited data. The shared task was focused on the classification of comments from a mental health forum, testing and escalating comments that require immediate attention to assist the forum's moderators. More about this event [47].

- 2016ko CLPsych: Also known as the Second Workshop on Computational Linguistics and Clinical Psychology. Same as the other workshops but with limited data. The shared task was focused on the classification of comments from a mental health forum, testing and escalating comments that require immediate attention to assist the forum's moderators. More about this event [48].

- 2015ko CLPsych: Also known as the Second Workshop on Computational Linguistics and Clinical Psychology. Same as the other workshops but with limited data. The shared task consisted on predicting the mental illness from data. More about this event [49].

Additional Corpus

In this section we describe the other data-sets that we encountered during the thesis and explain their origin, which is the composition, which are the labels, what type of information do they provide…In each data-set there should be a README file that clarifies any possible doubt about the use, structure and size of the data-set.

2012 Temporal Relations

The n2c2 data sets are provided as a community service [50]. They consist of fully de-identified clinical notes and products of challenges. They are freely available for the research community, but subject to a Data Use Agreement (DUA) that must be honoured. Each individual user must access the data independently through the DBMI Data Portal.

CLEF 2018

The challenge consists in performing a task on early risk detection of depression. The challenge consists of sequentially processing pieces of evidence and detect early traces of depression as soon as possible. The task is mainly concerned about evaluating Text Mining solutions and, thus, it concentrates on texts written in Social Media. The least bit of depression is noticed, but there is no level/class of depression. More about this event [44]. Apart from the timestamp, the data is saved in a .xml format and the structure is the following:

```
<INDIVIDUAL>

    <ID>subject16</ID>

        <WRITING>

                <TITLE> </TITLE>

                <DATE> 2017-02-20 14:54:10 </DATE>

                <INFO> reddit post </INFO>

                <TEXT>
                ...
                ...
                ...
                </TEXT>

        </WRITING>
</INDIVIDUAL>
```

It's made up of 10 sets, Task1 is divided in chunks [1-10]. It has two python processing programs and there are 820 different users each. The number of messages is described in writings-per-subject-all-test.txt file:

- Total of 544447 messages.
- Average of 664 messages for each user.

CLEF 2019

This data was used for predicting the depression risk [43]. The task consists of estimating the level of depression from a thread of user submissions. For each user, the participants will be given a history of postings and the participants will have to fill a standard depression questionnaire (based on the evidence found in the history of postings). The questionnaires are defined from Beck's Depression Inventory (BDI), which assesses the presence of feelings like sadness, pessimism, loss of energy, etc. The questionnaire has the following 21 questions:

```
1. Sadness
0. I do not feel sad.
1. I feel sad much of the time.
2. I am sad all the time.
3. I am so sad or unhappy that I can't stand it.

2. Pessimism
0. I am not discouraged about my future.
1. I feel more discouraged about my future than I used to be.
2. I do not expect things to work out for me.
3. I feel my future is hopeless and will only get worse.

3. Past Failure
0. I do not feel like a failure.
1. I have failed more than I should have.
2. As I look back, I see a lot of failures.
3. I feel I am a total failure as a person.

4. Loss of Pleasure
0. I get as much pleasure as I ever did from the things I enjoy.
1. I don't enjoy things as much as I used to.
2. I get very little pleasure from the things I used to enjoy.
3. I can't get any pleasure from the things I used to enjoy.

5. Guilty Feelings
0. I don't feel particularly guilty.
1. I feel guilty over many things I have done or should have done.
2. I feel quite guilty most of the time.
3. I feel guilty all of the time.

6. Punishment Feelings
0. I don't feel I am being punished.
1. I feel I may be punished.
2. I expect to be punished.
3. I feel I am being punished.

7. Self-Dislike
0. I feel the same about myself as ever.
1. I have lost confidence in myself.
2. I am disappointed in myself.
3. I dislike myself.

8. Self-Criticalness
0. I don't criticize or blame myself more than usual.
```

1. I am more critical of myself than I used to be.
2. I criticize myself for all of my faults.
3. I blame myself for everything bad that happens.

9. Suicidal Thoughts or Wishes
0. I don't have any thoughts of killing myself.
1. I have thoughts of killing myself, but I would not carry them out.
2. I would like to kill myself.
3. I would kill myself if I had the chance.

10. Crying
0. I don't cry anymore than I used to.
1. I cry more than I used to.
2. I cry over every little thing.
3. I feel like crying, but I can't.

11. Agitation
0. I am no more restless or wound up than usual.
1. I feel more restless or wound up than usual.
2. I am so restless or agitated that it's hard to stay still.
3. I am so restless or agitated that I have to keep moving or doing something.

12. Loss of Interest
0. I have not lost interest in other people or activities.
1. I am less interested in other people or things than before.
2. I have lost most of my interest in other people or things.
3. It's hard to get interested in anything.

13. Indecisiveness
0. I make decisions about as well as ever.
1. I find it more difficult to make decisions than usual.
2. I have much greater difficulty in making decisions than I used to.
3. I have trouble making any decisions.

14. Worthlessness
0. I do not feel I am worthless.
1. I don't consider myself as worthwhile and useful as I used to.
2. I feel more worthless as compared to other people.
3. I feel utterly worthless.

15. Loss of Energy
0. I have as much energy as ever.
1. I have less energy than I used to have.
2. I don't have enough energy to do very much.
3. I don't have enough energy to do anything.

16. Changes in Sleeping Pattern
0. I have not experienced any change in my sleeping pattern.
la. I sleep somewhat more than usual.
lb. I sleep somewhat less than usual.
2a. I sleep a lot more than usual.
2b. I sleep a Iot less than usual.
3a. I sleep most of the day.
3b. I wake up 1-2 hours early and can't get back to sleep.

17. Irritability
0. I am no more irritable than usual.
1. I am more irritable than usual.

```
2. I am much more irritable than usual.
3. I am irritable all the time.

18. Changes in Appetite
0. I have not experienced any change in my appetite.
la. My appetite is somewhat less than usual.
lb. My appetite is somewhat greater than usual.
2a. My appetite is much less than before.
2b. My appetite is much greater than usual.
3a. I have no appetite at all.
3b. I crave food all the time.

19. Concentration Difficulty
0. I can concentrate as well as ever.
1. I can't concentrate as well as usual.
2. It's hard to keep my mind on anything for very long.
3. I find I can't concentrate on anything.

20. Tiredness or Fatigue
0. I am no more tired or fatigued than usual.
1. I get more tired or fatigued more easily than usual.
2. I am too tired or fatigued to do a lot of the things I used to do.
3. I am too tired or fatigued to do most of the things I used to do.

21. Loss of Interest in Sex
0. I have not noticed any recent change in my interest in sex.
1. I am less interested in sex than I used to be.
2. I am much less interested in sex now.
3. I have lost interest in sex completely
```

Given a dataset with multiple users (for each user, his history of writings is provided) and it is needed to produce a file with the following structure: Each line has the username and 21 values. These values correspond with the responses to the questions above (the possible values are 0, 1a, 1b, 2a, 2b, 3a, 3b -for questions 16 and 18- and 0, 1, 2, 3 -for the rest of the questions-). The answers are in the following file *Depression Questionnaires_anon.txt*.

Evaluation should be based on:

- the overlapping between the questionnaire filled by the real user and the questionnaire filled by the system (number of correct responses).
- the absolute difference between the levels of depression obtained from both questionnaires (level of depression obtained from the real questionnaire vs level of depression obtained from the estimated questionnaire). The level of depression is simply obtained by summing the numeric values of the responses to the individual questions. This gives an integer value in the range 0-63.
- the depression level obtained from this questionnaire is regularly used to categorize users as: minimal depression (0-9), mild depression (10-18), moderate depression (19-29), and severe depression (30-63). A third method of evaluation will consist of assessing the systems in terms of how many users are correctly categorized (automatic questionnaire vs real questionnaire).

The structure of the data is similar to the one of CLEF 2018 it has a total of 20 user and the message quantity of each user differ.

MDDL

Its official name is "Dataset for Depression Detection via Harvesting Social Media: A A Multimodal Dictionary Learning Solution" [51].To make depression detection via social media, we need to get a batch of well-labeled data to train the models. We employed heuristical rule-based methods to construct two benchmark well-labeled depression and non-depression datasets on Twitter, which has mature APIs and is prevalent around the world. We obtain the personal information on social media and a piece of anchor tweet for one user. Besides, as people should be observed for a period of time according to clinical experience, all the other tweets published within one month from the anchor tweet are also obtained.

There are three sub-data-sets, D [1-2-3]. Although D1 and D2 are well labelled, D1 user depressions are low. So they built a bigger D3.

- **D1 Depressive Database**, named "positive". Based on the tweets between 2009 and 2016, we constructed a depression data-set D1, where users were labelled as depressed if their anchor tweets satisfied the strict pattern "(I'm/ I was/ I am/ I've been) diagnosed depression".
- **D2 Non - Depressive Database**. Named "negative", constructed a non-depression data-set D2, where users were labelled as non-depressed if they had never posted any tweet containing the character string "depress". We select the tweets on December 2016.
- **D3 Depressive Candidate Database**, named "unlabelled". Based on the tweets on December 2016, we constructed an unlabelled depression-candidate data-set D3, where users were obtained if their anchor tweets loosely contained the character string "depress". Although the depression-candidate data-set contained much noise, it contained more depressed users than randomly sampling.

Each database has a "tweet" folder with anchor tweets, a user "users" folder with personal information of Twitter users, and a "timeline" folder with tweets one month before the anchor tweet.Data is saved in .json files, with the following structure:

```
{"created_at":"Mon Apr 25 04:52:21 +0000 2016","id":724461019981053952,
 "id_str":"724461019981053952","text":"hi",...,"user":{"id":13270702,
"id_str":"13270702","name":"Zoë","screen_name":"zoesunderground",...}...}
```

The size of each data-set is the following(in Bytes):

- D1 1.8GB
- D2 16.9GB
- D3 202.4GB

Recovering Patient Journeys: A Corpus of Biomedical Entities and Relations on Twitter (BEAR)

Social media data, such as posts by patients and their relatives, provides valuable insights into the patient's perspective and journey with a medical condition, including subjective experiences, self-treatment, and self-diagnoses. The publicly available dataset consists of 2,100 tweets with approximately 6,000 entity annotations and 3,000 relation annotations. MOre information about this in [52].

Resources for automatic fact-checking in biomedical tweets

Resources for automatic fact-checking in biomedical tweets from University of Stuttgart [53]. Tweets with biomedical claims (BioClaim), tweets with annotated biomedical entities and relations (BEAR), and tweets with verdicts and evidence texts for fact-checking Covid-19 claims (CoVERT).

Bibliography

[1] Elizabeth D. Liddy. Natural language processing., 2001. See page 1.

[2] Unicode. Full emoji list, v15.0, 1991. See page 2.

[3] Michael M. Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893, 2019. See pages 4, 12, 31, 32, and 34.

[4] Mandar Deshpande and Vignesh Rao. Depression detection using emotion artificial intelligence. In *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, pages 858–862, 2017. See pages 4, 34.

[5] Kate Niederhoffer, Kristy Hollingshead, Philip Resnik, Rebecca Resnik, and Kate Loveys, editors. *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. See page 12.

[6] Nazli Goharian, Philip Resnik, Andrew Yates, Molly Ireland, Kate Niederhoffer, and Rebecca Resnik, editors. *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, Online, June 2021. Association for Computational Linguistics. See page 12.

[7] Farig Sadeque, Dongfang Xu, and Steven Bethard. Measuring the latency of depression detection in social media. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 495–503, New York, NY, USA, 2018. Association for Computing Machinery. See page 12.

[8] Hannah Burkhardt, Michael Pullmann, Thomas Hull, Patricia Areán, and Trevor Cohen. Comparing emotion feature extraction approaches for predicting depression and anxiety. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 105–115, Seattle, USA, July 2022. Association for Computational Linguistics. See pages 13, 31.

[9] Rohit Voleti, Stephanie Woolridge, Julie M. Liss, Melissa Milanovic, Christopher R. Bowie, and Visar Berisha. Objective assessment of social skills using automated language analysis for identification of schizophrenia and bipolar disorder, 2019. See pages 13, 32.

[10] Adrian Benton, Margaret Mitchell, and Dirk Hovy. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain, April 2017. Association for Computational Linguistics. See pages 13, 32.

[11] Adrian Benton, Margaret Mitchell, and Dirk Hovy. Multi-task learning for mental health using social media text, 2017. See pages 13, 32, 34, and 60.

[12] CLPsych. Clpsych 2021 shared task, 2021. See page 17.

[13] Susan Wang. Depression detection using twitter data - group project for udacity private and secure ai project showcase, 2019. See pages 17, 69.

[14] CLPsych. Clpsych 2021 shared task practice dataset, 2021. See pages 17, 20, and 69.

[15] Política Social e Igualdad. Axencia de Avaliación de Tecnoloxías Sanitarias de Galicia (avalia-t); 2010. Guías de Práctica Clínica en el SNS: avalia-t Nº 2010/02. Grupo de Trabajo de la Guía de Práctica Clínica de Prevención y Tratamiento de la Conducta Suicida. Guía de Práctica Clínica de Prevención y Tratamiento de la Conducta Suicida. Madrid: Ministerio de Sanidad. *La conducta suicida.* Axencia de Avaliación de Tecnoloxías Sanitarias de Galicia (Avalia-t), 2010. See page 18.

[16] University of Maryland. The university of maryland reddit suicidality dataset, version 2, 2019. See page 25.

[17] Jason. Full reddit submission corpus (2006 thru august 2015), 2015. See page 25.

[18] Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA, June 2018. Association for Computational Linguistics. See pages 25, 30, and 69.

[19] Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, June 2019. See page 25.

[20] Darcy J. Corbitt-Hall, Jami M. Gauthier, Margaret T. Davis, and Tracy K. Witte. College students' responses to suicidal content on social networking sites: An examination using a simulated facebook newsfeed. *Suicide and Life-Threatening Behavior*, 46(5):609–624, 2016. See page 29.

[21] Thomas E. Joiner, Gerald I. Metalsky, Jennifer Katz, and Steven R. H. Beach. Scientizing and routinizing the assessment of suicidality in outpatient practice. *Professional Psychology: Research and Practice*, 30:447–453, 1999. See page 29.

[22] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537, 2011. See pages 31, 68.

[23] Anjali Mittu. The baseline code for the clpsych 2021 shared task., 2021. See pages 31, 69.

[24] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, USA, 2014. See pages 32, 58.

[25] S.L. Gortmaker, David Hosmer, and S. Lemeshow. Applied logistic regression. *Contemp Sociol.*, 23, 01 2013. See page 32.

[26] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, jun 2008. See pages 32, 33.

[27] Chih-Jen Lin and Jorge J. Moré. Newton's method for large bound-constrained optimization problems. *SIAM Journal on Optimization*, 9(4):1100–1127, 1999. See page 33.

[28] Mohammad Hossin and Md Nasir Sulaiman. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1, 2015. See page 34.

[29] Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and f-score. pages 345–359, 01 2005. See pages 34, 35.

[30] Jin Huang and C.X. Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299–310, 2005. See page 36.

[31] Kazi Saeed Alam, Shovan Bhowmik, and Priyo Ranjan Kundu Prosun. Cyberbullying detection: An ensemble based machine learning approach. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pages 710–715, 2021. See page 36.

[32] Gary King and Langche Zeng. Logistic regression in rare events data. *Political Analysis*, 9:137–163, Spring 2001. See page 48.

[33] Elisa Mena-Maldonado, Rocío Cañamares, Pablo Castells, Yongli Ren, and Mark Sanderson. Popularity bias in false-positive metrics for recommender systems evaluation. *ACM Trans. Inf. Syst.*, 39(3), may 2021. See page 59.

[34] Foster Provost and Tom Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42:203–231, 01 2001. See page 59.

[35] Vankayala Tejaswini, Korra Sathya Babu, and Bibhudatta Sahoo. Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, nov 2022. Just Accepted. See page 60.

[36] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. See page 68.

[37] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, January 2007. See page 68.

[38] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification, 2018. See page 68.

[39] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Feb. 2015. See page 68.

[40] Kamaldeep Singh. How to improve class imbalance using class weights in machine learning?, 2020. See page 69.

[41] Scott Pakin. The comprehensive latex symbol list, 2008. See page 69.

[42] The Workshop on Computational Linguistics and Clinical Psychology. Clpsych 2022 shared task, 2014. See page 70.

[43] Javier Parapar David E. Losada, Fabio Crestani. CLEF eRisk: Early risk prediction on the Internet | CLEF 2019 workshop — early.irlab.org. https://early.irlab.org/2019/index.html, 2019. [Accessed 31-08-2023]. See pages 70, 72.

[44] Javier Parapar David E. Losada, Fabio Crestani. CLEF eRisk: Early risk prediction on the Internet | CLEF 2018 workshop — early.irlab.org. https://early.irlab.org/2018/index.html, 2018. [Accessed 31-08-2023]. See pages 70, 71.

[45] Annotation examples - brat rapid annotation tool. https://brat.nlplab.org/examples.html. [Accessed 31-08-2023]. See page 70.

[46] Kate Loveys, Kate Niederhoffer, Emily Prud'hommeaux, Rebecca Resnik, and Philip Resnik, editors. *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, New Orleans, LA, June 2018. Association for Computational Linguistics. See page 70.

[47] Kristy Hollingshead, Molly E. Ireland, and Kate Loveys, editors. *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, Vancouver, BC, August 2017. Association for Computational Linguistics. See page 70.

[48] Kristy Hollingshead and Lyle Ungar, editors. *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, San Diego, CA, USA, June 2016. Association for Computational Linguistics. See page 70.

[49] *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Denver, Colorado, June 5 2015. Association for Computational Linguistics. See page 70.

[50] Harvard and George Manson University. Data sets | national nlp clinical challenges (n2c2). https://n2c2.dbmi.hms.harvard.edu/data-sets. (Accessed on 08/31/2023). See page 71.

[51] Shen Guangyao. Github - sunlightsgy/mddl: Dataset for "depression detection via harvesting social media: A multimodal dictionary learning solution" in ijcai 17. https://github.com/sunlightsgy/MDDL. (Accessed on 08/31/2023). See page 75.

[52] Amelie Wührl and Roman Klinger. Recovering patient journeys: A corpus of biomedical entities and relations on Twitter (BEAR). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4439–4450, Marseille, France, June 2022. European Language Resources Association. See page 76.

[53] Institute for Natural Language Processing University of Stuttgart. Resources for automatic fact-checking in biomedical tweets | institute for natural language processing | university of stuttgart. https://www.ims.uni-stuttgart.de/en/research/resources/corpora/bioclaim/. (Accessed on 08/31/2023). See page 76.