A Data Driven Framework for Mental Health States Assessment in Social Platforms

Diana Ramírez Cifuentes

DOCTORAL THESIS UPF / 2022

Directors of the thesis:

Prof. Dr. Ricardo Baeza-Yates
Department of Information and Communication Technologies
Universitat Pompeu Fabra

Prof. Dr. Ana Freire
Department of Operations, Technology and Science
UPF Barcelona School of Management

DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGIES

Universitat Pompeu Fabra Barcelona

Dedicated to people struggling with mental disorders

ACKNOWLEDGEMENTS

This thesis has been completed thanks to the collaboration of many people and institutions. Therefore, I would like to thank all the people who have been part of this journey.

First, I want to acknowledge my thesis advisors, Ana Freire and Ricardo Baeza-Yates. I will always be thankful for their patience and support, and for having shared with me their time, experience and knowledge through these years.

Carrying out this thesis has been a learning experience thanks to its multidisciplinary nature. I am grateful for the collaboration of clinicians of the Department of Mental Health - Centro de Investigación Biomédica en Red de Salud Mental - Parc Tauli University Hospital (Sabadell): Nadia Sanz Lamora, Aida Álvarez, Alexandre González-Rodríguez, and Joaquim Puntí Vidal; members of the Fundación Instituto de Trastornos Alimentarios (FITA): Meritxell Lozano, Raquel Linares and Roger Llobet Vives; and to Pilar Medina, psychologist and member of the Communication Department at UPF.

I am also grateful for having collaborated with members of the Computer Vision Center of the Universitat Autonoma de Barcelona: Diego Alejandro Velazquez, Josep Maria Gonfaus, and Jordi Gonzàlez. I also had a great time collaborating with Esteban Ríssola (Università della Svizzera italiana).

One of the most enriching experiences of this journey was my research stay at the Hubert Curien Laboratory in Saint Étienne (University of Lyon). I am especially grateful to the guide, hospitality and support of Christine Largeron. It has also been a great experience to collaborate with researchers such as Julien Tissier and Mathias Géry.

I would also like to thank the members of the Information and Communication Technologies Department of Universitat Pompeu Fabra, including Lydia García, Ruth Temporal, Jana Safrankova, and Aurelio Ruiz. I would like to especially thank my colleagues, collaborators and ex-members of the Web Science and Social Computing Research Group (WSSC): Carlos Castillo, leader of the WSSC group; Maria Rauschenberger, Lorena Recalde, Eduardo Graells-Garrido, Meike Zehlike, David Solans, Francesco Fabbri, Valerio Lorini,

Fedor Vitiugin, Marzieh Karimi-haghighi, Manuel Portela, Paula Fortuna and Marina Estevez. Special thanks to Ana Freire, yet again, I cannot be more thankful for your support. Thank you as well to all the members of the Volleybolud@s beach volley team, including Silvia, Pablo, Adrián, and Rasoul.

I would also like to acknowledge the support and encouragement provided by my former teacher at the Escuela Politécnica Nacional (EPN), María Hallo. Thanks as well to my friends from the EPN: Marcelo, and Vanessa.

Finally, and most importantly, none of this work could have been possible without the support of my family and friends. First I would like to thank my parents: Jessica and Marcelo, thank you for all your unconditional love, for being my greatest support and for all your efforts. Thanks to my sisters: Nicolle and Mafer; thanks to my grandparents: Guillermo, Celina and Cecilia, and to my aunt and uncles: Taty, Guido, Rafael and Juan Carlos. Thanks to my family in Barcelona: Marcia, Arthur and especially Priscila, I cannot be more grateful to you, thanks for always being there for me. Special thanks to my lifelong friends: Daniela, Keila, Regina and Pamela; and to the friends I made through this journey: Marcela, Gabriel, Joaquín, Poleth, Xavier, Romina, Jessica, and Carlos.

Funding

This doctoral project has been supported by the Spanish Ministry of Economy and Competitiveness under the María de Maeztu Units of Excellence Program (MDM-2015-0502).

Research on this thesis has also received complementary support from the University of Lyon–IDEXLYON.

The funders did not have a role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ABSTRACT

The link between mental disorders and social media usage has led researchers to work on the development of automated methods to detect mental health issues in social platforms.

This thesis proposes a framework for mental health states assessment, considering suicidal ideation, eating disorders, depression and alcoholism as use cases. This framework is composed of modules dedicated to the characterization and detection of mental disorders, along with the definition of a non-invasive support-provision approach based on social recommendation.

We make use of user characterization techniques, and propose several predictive models based on behavioral and multimodal data. We also propose a contact recommendation approach, evaluated by people with anorexia nervosa, which prioritizes the recommendation of harmless accounts to follow in social platforms.

The main contributions of this work are: 1) insights regarding the behavior of users with mental disorders in social media; 2) methods for the enhancement of text representations (word embeddings) adapted to binary and multiclass predictive tasks that address domain specific tasks, and that handle small data; 3) several predictive models for the detection of mental disorders; and 4) the definition and evaluation of a contact recommendation method dedicated to users with anorexia. This method has been proven to be helpful for counteracting the over-personalization effects caused by social platforms' recommender systems.

This research work is focused on the analysis of data from Spanish and English speakers, addressing multiple social platforms. With the outcomes of this thesis we expect to contribute to the further development of tools to assist experts and help users living with mental disorders.

RESUM

El vincle entre trastorns mentals i l'ús de xarxes socials ha portat als investigadors a treballar en el desenvolupament de mètodes automatitzats per a detectar problemes de salut mental en xarxes socials.

Aquesta tesi proposa una estructura per a l'avaluació d'estats de salut mental d'usuaris, considerant com a casos d'ús la ideació suïcida, els trastorns alimentaris, la depressió i l'alcoholisme. L'estructura està composta per mòduls dedicats a la caracterització i detecció de trastorns mentals, juntament amb la definició d'un enfocament de provisió de suport no invasiu basat en la recomanació social.

Fent ús de tècniques de caracterització d'usuaris, proposem diversos models predictius basats en dades multimodals i de comportament. També proposem un enfoc de recomanació de contactes, avaluat per persones amb Anorèxia Nerviosa, que prioritza la recomanació de comptes inofensius a seguir en xarxes socials.

Les contribucions principals d'aquest treball són: 1) l'adquisició de coneixements sobre el comportament dels usuaris amb trastorns mentals en les xarxes socials; 2) mètodes per a la millora de representacions de text (Word embeddings) adaptats a tasques predictives binàries i multiclasse que aborden tasques específiques d'un domini i que no manegen dades massives; 3) diversos models predictius per a la detecció de trastorns mentals; i 4) la definició i avaluació d'un mètode de recomanació de contactes dedicat a usuaris amb anorèxia. S'ha demostrat que aquest mètode és útil com a manera de contrarestar els efectes de sobrepersonalització causats pels sistemes de recomanació de les plataformes socials.

Aquest treball de recerca es centra en l'anàlisi de dades de parlants d'espanyol i anglès, abordant múltiples plataformes socials. Amb els resultats d'aquesta tesi, esperem contribuir al desenvolupament de futures eines per a assistir a experts i ajudar als usuaris que viuen amb trastorns mentals.

RESUMEN

El vínculo entre trastornos mentales y el uso de redes sociales ha llevado a los investigadores a trabajar en el desarrollo de métodos automatizados para detectar problemas de salud mental en redes sociales.

Esta tesis propone una estructura para la evaluación de estados de salud mental de usuarios, considerando como casos de uso la ideación suicida, los trastornos alimentarios, la depresión y el alcoholismo. La estructura está compuesta por módulos dedicados a la caracterización y detección de trastornos mentales, junto con la definición de un enfoque de provisión de soporte no invasivo basado en la recomendación social.

Hacemos uso de técnicas de caracterización de usuarios, y proponemos varios modelos predictivos basados en datos multimodales y de comportamiento. También proponemos un enfoque de recomendación de contactos, evaluado por personas con anorexia nerviosa, que prioriza la recomendación de cuentas inofensivas a seguir en redes.

Las contribuciones principales de este trabajo son: 1) la adquisición de conocimientos sobre el comportamiento de los usuarios con trastornos mentales en las redes sociales; 2) métodos para la mejora de representaciones de texto (Word embeddings) adaptados a tareas predictivas binarias y multiclase que abordan tareas específicas de un dominio y que no manejan datos masivos; 3) varios modelos predictivos para la detección de trastornos mentales; y 4) la definición y evaluación de un método de recomendación de contactos dedicado a usuarios con anorexia. Se ha demostrado que este método es útil como forma de contrarrestar los efectos de sobrepersonalización causados por los sistemas de recomendación de las plataformas sociales.

Este trabajo se centra en el análisis de datos de hablantes de español e inglés, abordando múltiples plataformas sociales. Con los resultados esperamos contribuir al desarrollo de futuras herramientas para asistir a expertos y ayudar a los usuarios que viven con trastornos mentales.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Motivation

According to the World Health Organization (WHO) [142], mental and behavioral disorders are clinically significant conditions involving alterations in thinking, mood (emotions) or behavior. These conditions are linked with distress and/or problems functioning in social, work or family activities.

Since their appearance, online social networks (OSNs) have had a considerable impact in the way people communicate and interact. Studies have been done regarding the association of online social networks with certain mental disorders including depression, anxiety, bipolarity, eating disorders (EDs) and stress, among others [111].

Social platforms have eased the access to information that can provide negative feedback to people suffering from mental disorders, and has allowed users to share content promoting self-harming behaviors [155]. In contrast with this, people with mental disorders have found support in communities that promote pro-recovery content. An instance of this are pro-recovery communities for people living with eating disorders [163]. In fact, prior studies have stated that symptoms associated with mental disorders can be observable in online social networks and web forums [63,113].

Out of the social platforms' context, most of the screening tasks are performed by clinicians through in-person interactions with the usage of structured interviews and rating

scales. These interactions can only take place after a contact has been established between patients and clinicians. However, seeking help is not a simple task due to the stigma related to mental disorders, the lack of economic resources, and the difficulties given by the availability of appropriate healthcare facilities [146]. An instance of this is the US, where nearly 136 million people live in areas with a shortage of mental health providers [160]. This is how social platforms, which have a broader reach, can be used as a means through which clinicians can be put in contact with potential patients [134].

Within the computer science field, as stated by Guntuku *et al's.* review [63], the link between mental disorders and social media usage has led researchers to work on the development of automated methods to either detect mental disorders like depression, or their signs and symptoms such as suicidal ideation. These have been denoted as mental health status [25], or mental states [134] assessment approaches. In order to do so, certain characterization methods have been developed or adapted based on the analysis of data that users generate online [33,34,121].

The facts that motivate the current work are the following:

- The largest amount of research studies published regarding user characterization methods have been conducted primarily with the purpose of modeling user behaviors for marketing, advertising, subscription, membership and security perspectives [65,79,95]. However, the characterization of mental disorders in online social networks is a field with plenty of aspects waiting to be explored regarding the detection of symptoms, the interactions between users living with these disorders, and the changes in the behavior, vocabulary usage and topics of interest as they move towards recovery [63,69].
- Work has been dedicated to the development of predictive techniques for mental health states assessment online. Such work has been mainly centered in the usage of text based features, which have proved to be effective [63,134], however more accurate text representation techniques can be developed or adapted to the domain of mental disorders. Also, the analysis of images, and behavioral and relational data can provide further information [80].

- Proposals are required for the definition of non-intrusive support provision approaches once cases of risk are detected. This is relevant considering the impact that may cause in users a notification regarding their mental health status [25].
- Additionally, very few studies have assessed the analysis and development of predictive tools addressing Spanish speakers [49,87,90]. Among 75 studies reviewed in [26] between 2013 and 2018, only one study had used data in Spanish.

Working in these aspects can lead to the development of accurate predictive models based on features that characterize users with mental disorders, their conditions, and the role of their network within the social platform's environment. It can become a key factor for the development of early risk detection and contact recommender systems, which can evolve into tools capable of assisting experts on the diagnosis and treatment of mental disorders.

In this context the STOP project[1] surged, and it is dedicated to the prevention of suicide in online social platforms. The research work performed in this thesis is part of this project.

1.2 Goals

The main goal of our work is to develop a framework, understood as a conceptual structure, for mental health state assessment in online social platforms. The framework consists of elements dedicated to: 1) the characterization of users with mental disorders; 2) the definition of predictive models dedicated to the mental state assessment of such users; and 3) social recommendation approaches defined as nonintrusive support-provision methods, which may encourage users to seek for help.

We define as secondary goals the following:

- To identify the main features that characterize mental health states (in particular anorexia, and suicidal ideation) based on content (images and text), and behavioral data extracted from social media.

[1] https://stop-project.upf.edu/

- To develop new and effective automated methods for the detection of signs and symptoms of mental disorders in social platforms' users.
- To evaluate a noninvasive support-provision method based on the development of a contact recommender system dedicated to users with anorexia nervosa (AN).

The research questions we address in this work are the following:
RQ1) which are the textual, visual, relational and behavioral elements that characterize disorders such as depression, suicidal ideation, alcoholism and eating disorders like anorexia in social platforms?
RQ2) how can the features that characterize mental disorders be exploited for the development of new automated and explainable detection methods that can assist specialists to reach out to people at risk?
RQ3) Can a contact recommender system for users with anorexia nervosa connect them with pro-recovery communities so that users at risk are encouraged to seek help?

1.3 Challenges

The main challenges we face for mental health status assessment in online social platforms are the following:

- The data provided by social networks is limited for screening compared to the data that clinicians can gather through in-person interaction with patients. Even data related to simple demographic information (age and gender) from users must be inferred because it is not often disclosed by social platforms [166].
- To get trustworthy data we require the intervention of specialized clinicians as annotators, who due to time constraints cannot annotate large amounts of data. This means that the predictive models proposed should take into account the issues of working with small data [14].
- This is a multidisciplinary work. To design the methods proposed we require the intervention of clinicians and the participation of patients for evaluation purposes.
- Data shared in social media has different formats and thus we have to consider behavioral, textual and image-based

data in a way such that data of different types can be transformed into quantitative, measurable units.

- It is a challenge to propose a noninvasive way to provide assistance once cases of risk have been detected through predictive methods. This is mainly because of privacy, ethical [25] and even philosophical issues [32].

1.4 Contributions

All the material that makes part of this thesis has been taken from journal and conference papers that have been published during the course of the PhD research.

The main publications that describe our contributions are the following:

1. Ramírez-Cifuentes D, Freire A. UPF's Participation at the CLEF eRisk 2018: early risk prediction on the Internet. In. Cappellato L, Ferro N, Nie JY, Soulier L, editors. Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum; Avignon, France. CEUR Workshop Proceedings; 2018. P. 1-12. (described in Chapters 5, and 7 [98])
2. Ramírez-Cifuentes D, Largeron C, Tissier J, Freire A, Baeza-Yates R. Enhanced word embeddings for Anorexia nervosa detection on social media. Lecture Notes in Computer Science; Springer; 2020. P. 404–417. (described in Chapters 3, 4, 5 and 7 [128])
3. Ramírez-Cifuentes D, Freire A, Baeza-Yates R, Puntí J, Medina-Bravo P, Velazquez DA, Gonfaus JM, Gonzàlez J. Detection of Suicidal ideation on social media: multimodal, relational, and behavioral analysis. Journal of Medical Internet Research; 2020; 22(7):e17758. (described in Chapters 3, 4, 5 and 7 [126])
4. Ramírez-Cifuentes D, Freire A, Baeza-Yates R, Lamora NS, Álvarez A, González-Rodríguez A, Rochel ML, Vives RL, Velazquez DA, Gonfaus JM, Gonzàlez J. Characterization of Anorexia nervosa on social media: textual, visual, relational, behavioral, and demographical analysis. Journal of Medical Internet Research. 2021; 23(7). (described in Chapters 3, 4 and 7 [125]).
5. Ramírez-Cifuentes D, Largeron C, Tissier J, Baeza-Yates R, Freire A. Enhanced word embedding variations for the

detection of substance abuse and mental health issues on social media writings. IEEE Access 2021; 9:130449–130471. (described in Chapters 3, 4, 5 and 7 [127]).

6. Ramírez-Cifuentes D; Baeza-Yates R; Lozano M.; Freire A. A contact recommender system for social media users with Anorexia nervosa. Submitted, 2022. (described in Chapters 3, 6 and 7 [124]).

Other publications that are related and addressed briefly in this thesis are the following:

7. Ramírez-Cifuentes D, Mayans M, Freire A. Early Risk Detection of Anorexia on Social Media. Internet Science; In: Bodrunova S, editor. Internet Science. 5th International Conference, INSCI 2018, Proceedings; 2018 Oct 24-26; St. Petersburg, Russia. Cham: Springer; 2018. p. 3-14. (LNCS; no. 11193. LNISA; no. 11193). (Mentioned in Chapters 2 and 5 [129])
8. Ríssola EA, Ramírez-Cifuentes D, Freire A, Crestani F. Suicide risk assessment on social media: USI-UPF at the CLPsych 2019 shared task. In: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology: 2019 Jun 6; Minneapolis, Minnesota, USA. Stroudsburg: ACL; 2019. p. 167–71. (Mentioned in Chapter 5 [136])
9. Solans D; Ramírez-Cifuentes D; Ríssola E; Freire A. Gender bias when using Artificial Intelligence to assess Anorexia nervosa on Social Media. Submitted, 2022. (Part of the contributions are described in Chapters 4 and 5 [149])

The main contributions of this research work are:

- *The characterization of mental disorders on online social platforms:* we characterize suicidal ideation (publications 3 and 5) and anorexia nervosa (publications 2, 4, 6 and 9) in Twitter users using behavioral and multimodal data. To the best of our knowledge, we have performed the first work of this type about suicidal ideation at a user level in Spanish speaking users (publication 3). We have also been the firsts to take into account the different stages towards recovery of AN according to the trans-theoretical model of behavior change (publication 4). We have also characterized eating disorders, depression, suicidal ideation and alcoholism in

Reddit through the analysis of textual data (publication 5). In this sense we have established a comparative analysis among the lexical features that characterize these conditions and identified the particular terms that are mostly used by users living with each of these disorders.

- *Predictive models for mental state assessment:* we develop several predictive models for the detection of depression (publications 1 and 5), AN (publication 1, 2, 6, 7 and 9), and suicidal ideation (publications 3, 5, and 8) in social platforms. This includes the definition of models that are based in enhanced representations of textual data (word embeddings) (publications 2 and 5), models that address multimodal data (publication 3); and early detection approaches (publications 1 and 7), which take into account the time it takes for a model to emit a decision. We consider binary tasks addressing mental disorders and control cases; and multiclass detection tasks for the detection of a given disorder among other related disorders.
- *A contact recommendation method*: we define a non-intrusive support-provision approach which consists in a contact recommender system dedicated to users with anorexia nervosa (publication 6). The recommendation approach prioritizes the suggestion of harmless accounts to follow. We also define an evaluation measure dedicated to estimate how precise yet harmless a recommender system is.

1.5 Thesis outline

The thesis is organized as follows. After this introduction we find the background chapter (Chapter 2), where we provide a description of mental disorders, and in particular of those that we consider as use cases for the thesis.

We also describe definitions regarding users' characterization, predictive methods and recommender systems, along with related work about the characterization and detection of mental disorders on social platforms.

In Chapter 3 we explain the overall thesis methodology including the structure of the framework proposed and the details of the experimental designs.

Chapter 4 describes our work on the characterization of mental disorders making emphasis in the analysis of textual, visual, and behavioral features. We explore deeply anorexia nervosa and suicidal ideation as use cases. We also establish a comparative analysis of the features that characterize related disorders and substance abuse conditions such as depression, suicidal ideation, eating disorders and alcoholism.

In Chapter 5 we focus on the predictive methods developed for mental health states assessment. We describe the proposal and evaluation of several methods developed for the detection of AN, suicidal ideation, and depression in social platforms. We also define a multiclass predictive task dedicated to the detection of cases of depression, suicidal ideation, eating disorders and alcoholism. As a complement to these methods we explore early detection techniques and we assess gender related biases in a predictive model for the detection of AN.

Chapter 6 covers the development and evaluation of a contact recommender system dedicated to users with anorexia nervosa. In this chapter we evaluate a recommendation model that has the goal of suggesting accounts that do not share content that is harmful for users with anorexia. For this purpose, we develop predictive models based on the features analyzed and the predictive models developed in Chapters 4 and 5. Finally, we end with a chapter dedicated to conclusions and future lines of research (Chapter 7).

Chapter 2

BACKGROUND

2.1 Introduction

In this chapter we describe the topics that are relevant for understanding the methods used to reach our objectives. We first provide an explanation regarding mental disorders, we then describe basic background topics that cover characterization, and predictive techniques, along with the basics of recommender systems. We also present a review of the state-of-the-art that is relevant to our work.

2.2 Mental disorders and mental health states

The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) defines a mental disorder as *a "*syndrome characterized by clinically significant disturbance in an individual's cognition, emotion regulation, or behavior that reflects a dysfunction in the psychological, biological, or developmental processes underlying mental functioning*"* [8].

Within the mental disorders we find major depressive disorder, bipolar disorder, generalized anxiety disorder, anorexia nervosa, bulimia nervosa; schizophrenia, post-traumatic stress disorder (PTSD), among others [8]. These are disorders characterized by signs and symptoms that can even lead to death if not treated.

Within the computer science community, work has been dedicated to predict the presence of mood and psychosocial disorders in online social platforms [63,134], as well as to

assess related symptomatology like self-harm [112] or suicidal ideation [108]. Thus, to refer to both: mental disorders and their related symptomatology, authors have adopted the terms *mental health status* or *mental state* [26,134]. In this sense, the use cases mainly studied in this thesis are anorexia nervosa and suicidal ideation, and in a minor extension: major depressive disorder (referred in this manuscript as depression); and alcohol use disorder (referred as alcoholism). We describe briefly these conditions in the following sections along with the trans-theoretical model of health behavior change (TTM) [122], which explains how people progress toward recovery.

2.1.1 Anorexia nervosa

It is an eating disorder characterized by restriction of energy intake which leads to a significantly low body weight. People with anorexia nervosa (AN) have an intense fear of gaining weight or becoming fat, and show a disturbance in the way their body weight or shape is perceived by themselves [8].

In online social platforms, communities that promote harmful behaviors related to eating disorders have been traced. People taking part of such communities are surrounded by content that promotes and rewards unhealthy behaviors [163].

2.1.2 Suicidal ideation

According to [110], suicidal ideation is a term that is used to refer to a range of contemplations, wishes, and preoccupations regarding death and suicide. However, as the authors mention, there is not yet an agreement regarding the definition of this term as some authors include suicide planning as part of the definition [110]. Moreover, there are related terms such as *suicidality* that is used to refer to the existence of suicidal thoughts, plans and suicide attempts [13]. Regardless of this, suicidality is also used as a synonym of suicidal ideation.

In addition, the DSM-5 [8] has introduced the Suicidal Behavior Disorder as a condition for further study, which is mainly intended to refer to individuals who have made a suicide attempt within the last 24 months.

Given the lack of an agreement in definitions, and to clarify, for the current work, when we refer to suicidal ideation we consider suicidal thoughts, plans, attempts, and in overall, suicide risk itself.

2.1.3 Depression

Major depressive disorder is characterized by a daily depressed mood which implies feelings of sadness, emptiness and hopelessness.

People with depression have a markedly diminished interest or pleasure in almost all daily activities, show signs of insomnia, fatigue, and feelings of worthlessness. They also present a diminished ability to concentrate and are likely to have recurrent thoughts of death [8].

2.1.4 Alcoholism

Alcohol use disorder is characterized by a problematic pattern of alcohol use such that it is often taken in large amounts or over a long period.

Alcoholism is characterized by a recurrent alcohol use resulting in a failure to fulfill obligations at work, home, or school [8].

2.1.5 Model of health behavior change

There are multiple behavioral change models, among them, one of the most used is the trans-theoretical model of health behavior change [41]. It is described as an integrative method for understanding how people progress toward adopting and maintaining healthy behaviors [122].

This model identifies the following six stages of change: 1) precontemplation, where the individual does not know that there is a problem or that a change is required in their life and, thus, does not seek help; 2) contemplation, where the person simultaneously considers and rejects the change, while being conscious of the existence of a problem; 3) preparation, where the individual starts to take small steps toward behavioral change, believing that it can lead to a healthier life; 4) action, in which the person has changed their behavior and intends to sustain it; 5) maintenance, a stage wherein the person has maintained the behavioral change for a considerable period

(>6 months); and 6) termination, a stage wherein the individual has no desire to return to their unhealthy behaviors.

It is important to state that relapse, which implies returning from the action or maintenance stages to an earlier stage, is likely.

2.2 A data-driven approach

We propose a data and problem driven framework, which is based in a process described by [143] as the data science process (Figure 2.1).

Briefly, this framework describes a process where first raw data is extracted from the real world; in our case it would be data extracted from online social platforms in form of text and images shared by users, along with metadata provided by the platform (*e.g.,* posting times, times people liked a post, times the post was shared, etc.).

Later, this data is processed and cleaned, meaning that some data are transformed into the required formats. For instance texts are represented by the frequency of appearance of the terms they are composed by. Also, unnecessary or incomplete data may be removed.

Then, an exploratory data analysis is done so that the data can be understood better. In our use case, this step would provide most of the information required for the characterization of mental disorders, and for the further development of predictive models.

Later, we can perform predictive tasks using machine learning techniques and statistical methods. This step would be related to our use case with the development of predictive tools to detect mental disorders in social platforms' users.

Finally, the findings of the process can be interpreted and communicated with the usage of particular visualization techniques so that decisions based on the outputs of the whole process can be made.

Also, from the predictive tools developed we can go to a step that involves building a data product as a solution for the issues identified in the data analysis. For our case, such solutions imply detection tools and contact recommender systems.

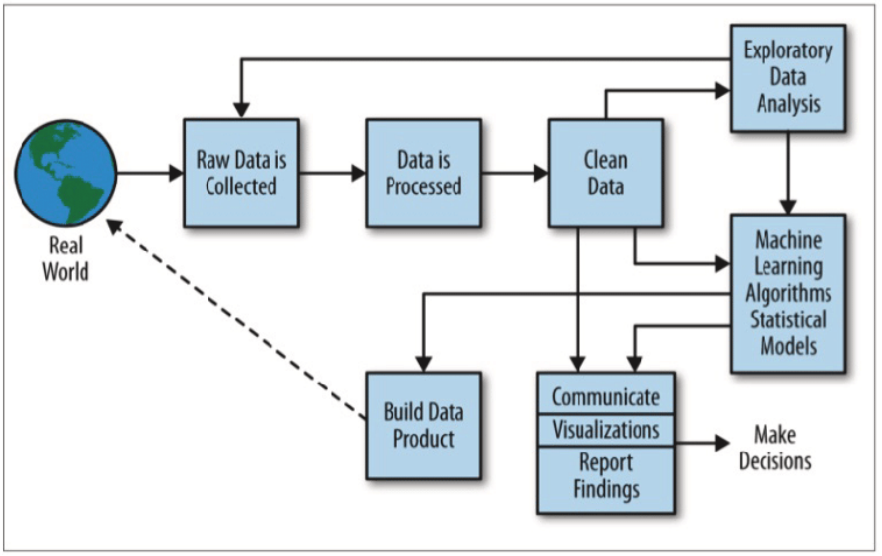

Figure 2.1 The data science process [143].

2.3 Characterization of users

Online social networks (OSNs) are defined according to the Oxford Handbook of Internet Studies [48] as networked communication platforms in which participants have uniquely identifiable profiles with content supplied by themselves, other users and/or system-level data. These sites can establish connections and are capable of consuming, producing, and/or interacting with content streams generated by users as part of their connections on the site.

OSNs are an object of interest for modeling users' behaviors. According to Tuna's *et al.* review [158], at first, this was key mainly for marketing, advertisement, and membership purposes. Nowadays the fields of application are numerous, going from finding hidden information from posted user data to identifying users' trustworthiness, mistrust, speciousness, and maliciousness in order to identify radicalization and civil unrest threats [60].

Characterizing users on OSNs involves collecting and analyzing users' data from different perspectives. The first one refers to identifying user's attributes such as age, gender, occupation and geo-location. This seems basic at first sight but it can become tricky taking into account that these features are not explicitly displayed in most social platforms. Knowing them

is important to recommend content and products based in the age and gender of people, among other applications.

Burger's *et al.* [23] findings, have shown that there is a difference in the way men and women make use of the language. Their method was based in the analysis of text features and the usage of neural networks for gender prediction. Deitrick *et al.* [42] have studied the usage of emoticons. Latest advances make use of deep learning techniques applied to text and images to infer the gender and age of users [166], whereas elements such as local dictionary words are defined to determine the location of a user [30,171].

The second perspective to take into account is the analysis of the behavior of a user. Different studies have proved that behavioral features can be of importance for the prediction of personality. For this task, features based on the frequency and intensity of interactions along with the priority and reciprocity are explored [58]. This perspective involves the identification of deceptive behaviors [150], and radicalism detection [2].

The third perspective to consider, as described by [158] are mental models, which refer to how the human mind represents all types of situations. These are of importance for this thesis in terms of the language usage [153], opinion and interest analysis. The work done regarding the analysis of the posts of a user in different time periods, provides an insight on interests and opinion changes [61]. Most of the work done regarding this field, has been applied to model social-issues and political preferences as part of a wider type of tasks dedicated to stance detection [6].

The last perspective to take into account for user characterization is user categorization. This is mostly oriented to identifying spammers, bots, fake users, and for entity resolution. These techniques mainly focus on the analysis of the relationships between users, posts content and frequencies, as well as the account properties [158].

The latest publications regarding user characterization, modeling or profiling methods have introduced the usage of word embeddings for text analysis along with deep learning techniques [172]. Researchers have started to mind as well about the usage of multi-modal data, taking into account that users not only share written content but also images and videos [50].

As it will be described in the following sections, given the complexity of mental disorders, a combination of the methods applied to each perspective should be considered for the characterization of users with these conditions.

2.4 Predictive models

Predictive modeling makes use of statistics techniques to predict and forecast outcomes with the aid of existing and historical data [84]. It often analyzes current and historical data to make predictions by using techniques from statistics, data mining and artificial intelligence (AI).

Within AI, machine learning (ML) helps a computer model to adapt to new circumstances and to detect and extrapolate patterns by learning from data. A model learns as the performance measure of the model increases for a given task.

A ML model takes as input instances that are provided in terms of features, along with labels assigned to each instance representing the desired output. For instance, for a depression detection model, an instance would represent a user, which at the same time will be represented by features such as the number of self-references made in their texts, the average number of times their posts have been shared, the number of friends, among others.

The label for the user would be whether they are an instance of a case of depression or of a control case. Multiple instances are fed to the model so that it can be trained using different types of learning techniques to later predict if a new unseen user is likely to have depression.

The main learning approaches in machine learning are supervised and unsupervised learning. Supervised learning, as in our prior example, corresponds to the case where the model is offered a set of instances with their respective labels as input, and thus it learns to identify the correct label.

The most common types of supervised learning are classification and regression techniques. In our case, classification techniques are the ones that will be mainly used. For this case, the label assigned to an instance takes the name of a class, and thus the goal of the model is to identify if an instance corresponds to one or another class.

Unsupervised learning is when the labels for the inputs are not provided. Models learn similarities between the input instances and try to predict a label based on these similarities.

In this thesis we make use of algorithms for classification tasks. These algorithms provide functions that weight the input features so that the output separates one class from another. Among the main algorithms used we find Random Forest (RF) [22], Support Vector Machines (SVM) [40] and Logistic Regression (LR) [168]. We also evaluate deep learning methods such as Multilayer Perceptron (MLP) [97] and Convolutional Neural Networks (CNNs) [169], which are machine learning approaches based on neural networks.

There are some evaluation measures that are commonly used to evaluate predictive models, which output for each instance the class to which it is assigned. Such predictions can be right or wrong. Thus, for this thesis, the evaluation measures used are Precision (P), Recall (R), F1-Score; Area Under the curve (AUC), and Accuracy (A) [55].

2.5 Contact recommender systems

Within online social networks, the main goal of contact recommendation is to find users in a social network from which a given user would benefit from relating to [140], so that they can be presented as suggestions for the user to either befriend or follow. Such suggestions are interpreted as users with whom the user might want to engage in an online network.

Popular social platforms offer user recommendation services. Instances of this are the ‘Who-to-follow’ service on Twitter [64] or the ‘People you may know’ services on Facebook and LinkedIn.

The objective of a common contact recommendation model [85] is to rank on top the accounts that the user is more likely to follow or befriend, under the principle that people tend to connect with users who they are likely to know (users that are part of their social network) or that have interests in common (interested in similar contents).

The output of a recommender system is an ordered list of users ranked according to criteria defined by recommendation algorithms [85,140] for a given target user. The evaluation of such systems consists in calculating measures such as

Precision (P), Recall (R) and the Mean Average Precision (MAP) at a given number (k) of ranked users recommended.

Precision, tells us about the amount of recommended candidates that would have actually been followed by target users over the total amount of users recommended at k; recall, measures the ratio of candidates recommended that the target user would actually follow at k, over the number of all candidates that the target user would actually follow; and the mean average precision takes into account how well ranked by the model were the users that target users are willing to follow [15].

2.6 Characterization of users with mental disorders

This section provides an insight on the work that has been done until now regarding the characterization of users with mental disorders on OSNs and social media in general.

This literature review focuses mainly on mood and eating disorders taking into account the scope of the thesis.

Studies have traced mental disorders through the analysis of social media data [26,63,134]. Textual cues have proved to be the most relevant for characterizing mental disorders [134].

Mainly lexical features that analyze structural, syntactical, topical, linguistic style, and domain specific elements are defined in order to compare users with mental disorders and control users (users that do not show signs or symptoms of such disorders) [26,63,134].

The main findings of related work regarding the textual cues that characterize depression are a high use of first person pronouns, expression of negative emotions, use of references to antidepressants, and symptom-related words. There are also expressions of anger, anxiety, hopelessness, and suicidal thoughts [34,38,113,132,134,144].

Regarding eating disorders, it has been found that pro-anorexia and pro-recovery communities exhibit distinctive affective, social, cognitive, and linguistic style markers, as pro-anorexics express greater negative affect, feelings of social isolation, and self-harm [33].

Users with high suicide probability were characterized by a high usage of pronouns, a low usage of verbs and a greater word count compared to those with lower risk [29].

Regarding the analysis of the behavior and relationships among users. The work of De Choudhury *et al.* [34] on Twitter defined an egocentric social graph, where a relationship is given by the interaction (reply) between 2 users.

The measures obtained were given by network properties and counts of the interactions [34] meaning that the content shared or discussed was not analyzed, nor the profile of the users with which depressed users are related.

Regarding suicidal ideation, Masuda *et al.* [99] found that the number of communities to which a user belongs to, and the fraction of suicidal neighbors in the social network (homophily), contributed the most to suicidal ideation.

Colombo *et al.* [36] examine the connectivity and communication of suicidal users on Twitter, and poses a study of the suicidal community based on the analysis of retweets between suicidal users.

These studies provide a starting point regarding the relationships of users with suicidal intentions but it leaves aside the analysis of the content shared, the behavioral changes across time, and the relationships with “non-suicidal" users.

The work of Lin *et al.* [107] is one of the few works that performs a deep analysis of the interactions and the network of the users studied. This work analyzes stress and has found that users’ stress states are revealed by the structure of their social interactions, including structural diversity, social influence, and strong/weak ties.

There has also been work that studies behavioral elements that tell us about the activity of the user in the platform. Most of this information is given by the analysis of posting frequencies [26]. Also, the work of Vedula *et al.* [161] found reduced and nocturnal online activity patterns in depressed users in the US.

Other elements studied involve the analysis of images [62,89,131,145,165]. An instance of such studies [131] found that photos posted to Instagram by depressed individuals were more likely to be bluer, grayer, and darker.

There are also studies that include demographical data such as age, gender, education, income and relationship status [26].

2.7 Screening users with mental disorders

Predictive models are built to perform the automated analysis of social media. These models use features or variables that have been mainly extracted from labeled user-generated data [26,63,70,134]. To collect the data, participants are either recruited to take a survey and share their social network account data [34,132,157], or data is collected from public online sources like Twitter, Facebook or Reddit [11,16,19,38,76,121].

Regarding the features that are extracted to build predictive models, according to [26,134] the most common ones are language features (textual cues) such as features that describe the structural or syntactical composition of posts (length of the post, part of speech tagging, use of emoticons, etc.), character and word models (n-gram use, word embeddings, etc.), topical features (topic modeling such as Latent Dirichlet Allocation (LDA)), linguistic style (use of dictionaries, readability coherence and subjectivity measures, etc.), and domain specific features that imply the development of lexicons related to the specific disorder.

Other types of features are based in the user behavior, which include the activity of the user (posting frequencies), interaction features that involve the interactions between users (follower-followee relationships, mentions, replies, retweets, etc.), and network features that imply the analysis of the network or graph structures of an individual (clustering coefficients, strong and weak ties, homophily, etc.).

There are also features that measure emotion and cognition elements (sentiment and psycholinguistic features), features that evaluate demographic aspects (age, gender, income, etc.), and image features (types of colors used, brightness, saturation, number of faces detected, etc.).

The algorithms selected for predictive purposes are mainly machine learning and statistical modeling techniques. According to [26] the most common algorithms are SVM, LR, and RF. The latest work has included deep learning approaches [152] for which the volume of data managed is relevant considering the difficulties of obtaining trustworthy labeled data, which generally is low (<1000 user level instances) [39,92,146].

Regarding the evaluation of predictive systems, the most used evaluation measures are those dedicated to assess machine learning models such as Precision, Recall, F-Score, and AUC [26]. In this context, the work of Losada *et al*. [91] recalls the importance of early detection and thus proposes a new metric dedicated to evaluate how fast the detection is done, known as the early risk detection error metric (ERDE).

2.8 Noninvasive support provision

There have been interventions addressing mental disorders through Internet and mobile based interventions (IMIs), such approaches have proved to be highly efficacious when they are compared to untreated controls [46]. Most of these interventions imply the development of mobile applications [59] and interventions provided through web platforms like cognitive behavior therapy provision online [102].

Very few interventions or support provision approaches have been dedicated to social networks [133]. In this sense, new social platforms have been developed and evaluated targeting people with mental disorders as users, that is the case of the Moderated Online Social Therapy Model (MOST) [86], which was designed as a framework for first episode psychosis patients. The platform allows structured interactions to occur via a forum through which users can share coping strategies, and there is a group therapy feature for social problem solving.

Just in-in-time adaptive mechanisms [46] also represent an intervention alternative. These type of intervention aim to predict changes in an individual's status to deliver personalized support when a person needs it most. Regarding mental disorders, there have only existed frameworks [46] that do not consider the usage of data extracted from social media, but mainly analyze information provided either by the user through an app and data collected by phone and wearable devices.

Popular social platforms such as Instagram[2], Twitter[3] and Facebook[4] have opted for intervening by either asking users to report cases of users at risk, or by developing AI powered tools

[2] https://help.instagram.com/553490068054878/?helpref=hc_fnav
[3] https://help.twitter.com/en/safety-and-security/self-harm-and-suicide
[4] https://www.facebook.com/safety/wellbeing/suicideprevention/

to detect automatically cases of risk and for either case they would deliver messages to the users at risk with information regarding how to reach for help. In Instagram for instance, when search terms of risk are used, before showing the search results, users are warned regarding the triggering content, and are offered information regarding sites to seek for help.

Considering social recommendation, there are no contact recommendation approaches to assess the issue of systems suggesting harmful content to people with eating disorders. However, there has been intense work dedicated to the detection and mitigation of echo chambers and filter bubbles [31]. These approaches are intended to enhance the suggestion of content that reduces the polarization among communities.

2.9 Ethical issues

Related work has discussed the ethical implications of developing predictive tools for mental health state assessment [25]. There are concerns about the usage of data from social media users, as despite being information that users agree to share in the terms of service of social platforms, they are not aware of the usage purposes.

Health data represents sensitive information, which can be used to train models for their misusage. Guntuku *et al.* [63] mention the usage of depression detection tools by employers and insurance companies as instances of possible misusages of such tools. This means that there is a need for a legal framework capable of assessing the proper usage of such tools, which can be highly beneficial if used for the right purpose as mentioned by Chancellor *et al*. [25].

In addition to the privacy concerns and the possible misuse of the predictive tools, there are other elements to analyze before their release. It is relevant to assess aspects related with the design of predictive models regarding biases and fairness issues.

These models should also have a reliable data collection and annotation approach; they should be accurate enough; and they should provide valid and interpretable outputs, which can justify the decision of the algorithms used and can be understood by clinicians [25].

Most of the research work has been focused on characterization and detection tasks, and very few tools have been placed in production. Instances of tools that have been deployed are the app Samaritan's Radar, which scanned a person's friends for concerning Twitter posts; and Facebook's automated tools to identify individuals contemplating suicide or self-injury [25]. Samaritan's Radar was pulled from production because it collected data without permission and also it was enabling harassers to intervene when someone was vulnerable. Regarding Facebook, the company has not been able to deploy their AI powered suicide prevention tools in the European Union (EU) due to the EU's Data Protection Directive and General Data Protection Regulations [39].

With this background, most authors agree on the fact that further effort shall be done to overcome ethical issues prior to the deployment of predictive models and their derived interventions. They analyze the ethical benefits of e-Health: like a broader access to treatment, more options for communication between clinicians and patients, and potential cost savings; and the elements that should be assessed before the application of such interventions like privacy and confidentiality issues, identity verification, data validity, trustworthiness of the models (fairness and biases), the role of clinicians and machines, crisis intervention, and legal concerns [71].

Chapter 3

METHODOLOGY

3.1 Introduction

In this chapter we describe the framework proposed. We first describe the problem we address. Then we describe our proposal, which is given by the framework architecture and experimental setups. We also describe the data gathering processes for our experimental procedures, then the evaluation approaches, and finally the ethical assessment procedures.

3.2 Problem identification

Mental disorders when untreated can lead to severe health issues and death in the worst case [54]. Social platforms can be a means through which people with undiagnosed conditions can be reached. This can lead them to seek proper treatment in order to recover from their condition [134,146].

In terms of the characterization of mental disorders, taking our use cases as a sample (depression, anorexia nervosa, alcoholism and suicidal ideation), we observe that there is a lack of work regarding the analysis of the stages that users go through toward recovery from mental disorders. Also, further research is required in the analysis of images, and behavioral data, which includes the analysis of the relationships between users.

In addition, most of the work done has addressed depression [26,63,134], and although there are more elements to analyze regarding depression, there are other conditions that have not been explored with the same depth.

We can see that most of the work done has been focused in the development of predictive models based mainly in text cues [26,63,134], which have proved to be the most useful for these tasks, and thus further research in the improvement of text representations adapted to the domain of mental disorders are required. Also, to complement textual elements, more models that combine diverse modalities of data are required in the definition of predictive models. This also includes the need for characterization studies and models that address users speaking other languages [26].

It is also necessary to define new predictive approaches that address the detection of multiple disorders. We can also notice the rise in the usage of deep learning techniques [26,63] which for some cases have obtained improvements in accuracy. The issue with such methods relies on the explainability of the decisions made, and the volume of data addressed, which tends to be small. Here, it is also required to involve clinicians in the definition of features and the interpretation of the output of predictive methods.

Another aspect to take into account is that studies mostly address predictive tasks that do not take into account a real time detection context, where it is important to make an accurate decision on time, especially in suicidality detection. In this sense, early detection systems are an element of interest [92], as well as non-invasive interventions or support procedures that can assist users to seek in person or remote help to overcome mental disorders.

Above all, even though there's been work done for the detection of mental disorders, there are very few proposals on tools and procedures that can make use of such predictive models. To the extent of our knowledge, there are no studies that address a complete model or framework to assist users with mental disorders. Such a framework should assess characterization, detection and non-invasive assistance or support elements.

The methods that are mostly related to our proposal are just in-in-time adaptive interventions which are likely to make use

of predictive methods prior to the intervention, but there have not been approaches of this type reported in the social media context [37,46]. In fact, the area of interventions or support provision requires more research work to be done. Hence, contact recommendation is an alternative proposed and evaluated in this research work.

3.3 Proposal

We propose a framework to assist in the diagnosis and treatment of users with mental health issues. To build it, we intend to explore methods capable of characterizing users with mental disorders on social media so that appropriate predictive models can be built for their application on detection and recommender systems. The framework will be composed of different modules that will address 1) characterization, 2) detection and 3) contact recommendation elements.

We have defined an evaluation framework to address multiple mental health states such as anorexia nervosa, depression, suicidal ideation, and alcoholism. Anorexia was selected as the main disorder to evaluate considering all the modules in the framework, whereas the remaining conditions are studied in the characterization and detection modules.

3.3.1 Architecture of the framework

The framework is composed of three modules as can be seen in Figure 3.1. First, 1) different types of features are extracted from users' data in order 2) to perform a characterization analysis of the mental health state given. We define a set of generic feature types, which are based on language, behavioral, demographic and image elements. We also define features that are specific to the condition studied.

After the characterization of a given disorder, we 3) proceed to make use of either the same features analyzed or a set of selected features to develop predictive models. Finally, 4) we make use of the techniques developed for the detection of users at risk in order to provide personalized contact suggestions (users to follow or befriend) through a recommendation approach that modifies the objective function for ranking users in order to prioritize the suggestion of users that do not share harmful content and pro recovery accounts.

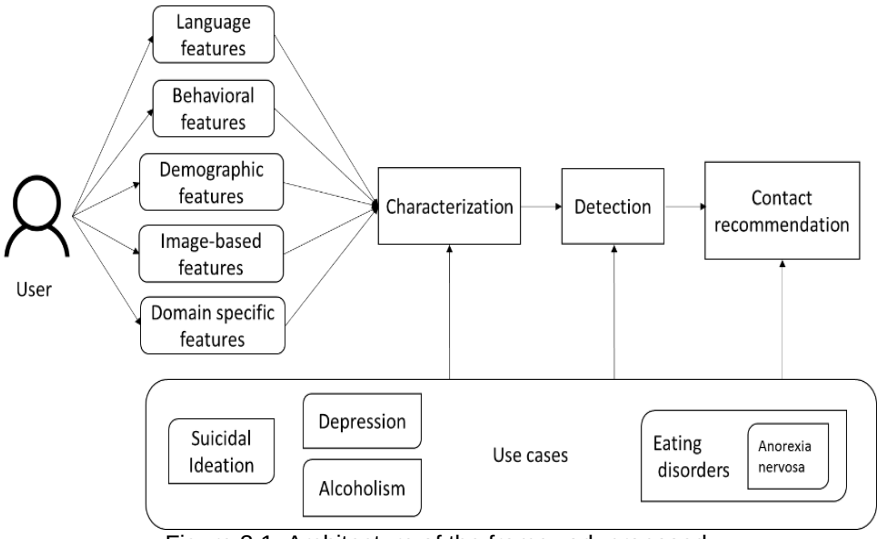

Figure 3.1. Architecture of the framework proposed.

We believe that this is a non-invasive way to provide assistance. It may lead target users to connect with people and specialized centers, and eventually to seek treatment.

Within our experimental framework we address suicidal ideation, depression, anorexia nervosa and alcoholism as use cases, being anorexia nervosa the disorder that is addressed through all the modules of our framework.

We provide details of the characteristics of every module, and the experiments performed in the following subsections.

3.3.2 Characterization module

This is a module in charge of identifying elements that characterize the mental health state of users.

Taking into account a data driven process [143], users' characterization would be compatible with the data collection, processing and cleaning; and with the exploratory data analysis step, where data is inspected and compared in order to identify particular patterns and characteristics that distinguish certain users' types from others.

Given a mental disorder to study, and an OSN, we collect data from multiple users that have been labeled by clinicians as likely to be showing signs and symptoms of the mental disorder that will be studied.

We also collect data from control users. The data collected from each user consists in a large sample of their posts, and metadata from the user's profile.

Once data has been collected, we extract features from it. The main feature types extracted and analyzed are the following [26,134]:

- *Language features:* these are features extracted from the text of the user's post. We analyze 1) structural, syntactical and linguistic style elements like the use of verbs, pronouns, length of the text, etc., mostly with the usage of lexicons like Linguistic Inquiry and Word Count (LIWC) [130]; 2) word models based in n-grams, which are representations of texts based in the frequency of the terms or sets of terms, and word embeddings where terms are represented in a vector space. We also analyze 3) features with topic modeling approaches and predefined dictionaries of topics like Empath [52].
- *Behavioral features:* These are features gathered for the analysis of 1) the users' activity, which takes into account posting frequencies at different scales of time (day/night, week/weekend, and seasons); 2) we analyze interactions between users through the analysis of topics of interest of the followees of users with mental disorders. We also analyze 3) the social network of users (graph structure) and detect communities among users at different stages of the disorder.
- *Emotion and cognition elements:* These are features that perform an analysis of sentiment, and psycholinguistic aspects. To obtain such features we make use of external tools as senti-py [73] and lexicons like Emolex [104] and LIWC [130].
- *Demographic features:* we infer the age and gender (male and female) of users. We also infer if an account belongs to an organization.
- *Image-based features:* We study images posted by users and their profile pictures. We analyze the outputs of detection models trained with images corresponding to the mental disorder studied and control images. We also evaluate properties of the images such as if an image is light or dark, if it is gray scale or not, if it has text, if it has faces

in it, along with the usage of tools for the detection of objects in images.

- *Domain-specific features:* These are features that are particular to the domain of the disorder studied. An instance is the usage of lexicons that describe specific signs and symptoms of a given disorder, such as the use of names of laxatives when characterizing anorexia.

Within this module we contribute to the state of the art by performing a comparative analysis of these features between mental disordered and control cases to identify features that distinguish risk cases from controls.

We also perform comparative analyses of features among the different stages of a mental disorder and among various disorders. For certain use cases we even define different control groups.

Through the analysis performed we have found interesting patterns and elements that characterize each of the conditions studied in English and Spanish speakers. Further details on our findings will be described in Chapter 4.

3.3.3 Detection module

Based on the characterization findings, we use the features analyzed to build and evaluate predictive models for mental health states assessment. In a data driven process, this module would match the machine learning algorithms and statistical models step [143].

Our main contributions within this module are the evaluation of several predictive models, the definition of a method to enhance the representation of textual elements, and the definition of models that address early risk detection approaches.

We also take into account issues related to fairness and gender in predictive models. We describe the details of our contributions regarding the detection module in Chapter 5.

3.3.4 Contact recommendation module

This module addresses what would be done once a predictive model detects signs and symptoms of a given mental disorder

in a user. This is a non-invasive support-provision approach addressed to social platforms.

According to the data driven process [143] this module would correspond to a data product based on the outcomes of the characterization and detection steps.

The approach proposed takes into account the characteristics of a platform like Twitter, where target users can decide who to follow among a list of users suggested by the platform.

The recommendation approach, targeted to users detected at risk, consists of modifying the objective function used by the recommender in order to place harmless recommendations (of accounts to follow/befriend) at the top of the ranking.

This solution is evaluated taking into account the anorexia use case, and it is inspired by the hypothesis that social platforms are likely to reinforce harmful behaviors in people with eating disorders, as they are likely to recommend like-minded users with common interests. Such interests are often harmful for the target user. We prove this hypothesis within this research work.

The details regarding our findings and the specific methods used for evaluating the recommendation method are detailed in Chapter 6.

3.4 Experimental methodology

Our experimental setup addresses characterization, detection and social recommendation methods applied to the use cases selected.

- Characterization

After collecting and annotating data through the usage of keywords related to it, we define feature types (Section 3.3.2) and perform comparative analyses between each use case and control cases [125–127], and between all the use cases [127]. In addition, taking anorexia as an instance, we perform an analysis of the stages toward recovery according to the trans-theoretical model of change behavior [125].

We investigated whether significant differences existed between the features of the instances of each use case or stage. We did this through the usage of statistical methods

such as the Mann Whitney U or chi-squared (X^2) tests, and correlation analyses [125–127].

Considering our use cases, we identify the main elements that characterize suicidal ideation using data in Spanish [126] and English [127], and different social platforms such as Reddit and Twitter.

Regarding anorexia, we perform a deeper analysis, which inspects the process toward recovery, making an emphasis in the analysis of the interest shared by AN users and their network [124,125]. For this use case we also analyze data in Spanish [124,125] and English [127], both from Reddit and Twitter.

In the case of depression and alcoholism, the characterization is mainly focused in the comparison of instances from both cases with other use cases, focusing on the analysis of text based features extracted from Reddit posts in English [127].

- Detection

We model detection problems (Section 3.3.3) as binary (use case vs. control) [98,126,128,129,136] and multiclass (multiple disorders) [127] supervised classification tasks. We build several models that combine features among those described in Section 3.3.2. Domain-specific features depend on the mental disorder for which the model is being built.

As classifiers, we use machine learning algorithms such as Logistic Regression, Random Forest, Support vector Machines, Multilayer perceptron and Convolutional neural networks.

We also address early risk detection tasks that face real time detection cases where it is required to incrementally build a representation of each user. In this sense we train a classifier that only emits a decision when it has enough information to output a positive decision with enough confidence [98,129]. We also handle the relevance of missing positive cases (False negatives) by assigning them a higher cost compared to false positive cases. This is important as missing cases of risk can imply an omission of the treatment required.

Finally, we use techniques to assess feature importance [126,149], and address gender biases in detection models by inspecting the features that are more relevant for the model and for clinicians for the detection of a disorder [149].

Taking into account the use cases addressed, for suicidal ideation we develop a multimodal predictive model to detect suicidal ideation cases and distinguish them from control cases. In this work we explore the contribution of different feature types in predictive models addressing data in data in Spanish [126] .

Regarding anorexia, we propose and evaluate early risk detection models that make use of text based features [98,129]. We also evaluate models that make use of enhanced word embeddings, adapted for the detection of anorexia [128]. Both of these approaches use Reddit datasets in English. We also create a model that detects AN users, who are at a stage when they are more likely to search for assistance. This model is useful for the definition of the recommender system proposal. Data for this case was collected from Spanish speaking users in Twitter [124]. We also develop models with the intention of assessing gender biases in them. This is done through the detection of the most predictive features in models trained only with data from male and female users [149].

For the case of depression, we create a predictive model for early risk detection [98] similar to the anorexia case. We did not focus strongly in developing new models for this use case in particular as is one of the most studied use cases in the state of the art [26].

Finally, we develop a multiclass predictive model for the detection of cases of eating disorders, depression, suicidal ideation and alcoholism. In this sense we extend the prior work dedicated to generate enhanced word embeddings so that it can also be used on multiclass predictive tasks [127].

- Contact recommendation

With data extracted from Twitter, and with the collaboration of specialists, and volunteers at the last stages of AN treatment, we design and evaluate an approach that modifies the objective function of a content and topology- based recommendation algorithm, in order to maximize the suggestion of harmless accounts for AN users [124].

The recommender system proposed is a product resulting from the characterization and detection steps. For the definition of this model we make use of content and topological features that are commonly defined for social recommendation. We also introduce a filtering method, and

what we denote as a harmlessness factor that penalizes harmful accounts in the suggestion rank.

In addition to developing a classifier that detects users with anorexia, we develop a classifier that distinguishes users that share harmful content from uses that share harmless content.

Our main contributions with this work are the architecture and the objective function of the recommender system, the predictive models defined, and a new evaluation measure to assess how harmless a recommender system is.

We summarize in Table 3.1 the use cases addressed, the data sources and languages analyzed, and the modules of the framework that have been explored for each use case.

Table 3.1. Summary of the mental disorders, data sources, languages and modules explored.

Mental disorder / state	Data sources	Languages	Module explored
Mental disorders (general)	Reddit	English	Characterization and detection
Eating disorders (general)	Reddit	English	Characterization, detection
Anorexia Nervosa	Reddit and Twitter	English and Spanish	Characterization, detection and contact recommendation
Suicidal ideation	Reddit and Twitter	English and Spanish	Characterization and detection
Depression	Reddit	English	Characterization and detection
Alcoholism	Reddit	English	Characterization and detection

3.4.1 Data collection and annotation process

We built Twitter datasets to assess suicidal ideation, and anorexia. We selected Twitter as our main data source, as it has been proven to be suitable for analyzing mental disorders on social media [1,11,24,39], including suicidal ideation, and eating disorders.

We also highlighted the following aspects that this platform offers for our research: 1) the possibility of having posts in multiple languages; 2) the chance of finding relational and behavioral factors; and 3) the provision of a set of chronologically ordered posts from each user.

We also built a dataset using data from Reddit to analyze and evaluate textual cues for the detection of alcoholism, eating disorders, depression and suicidal ideation.

The datasets defined are the following:

- *Dataset 1 – Depression and anorexia* [98,129]

To develop predictive approaches for the detection of depression [98] and anorexia [98,129] we used the dataset provided by Losada *et al.* [91,92]. The dataset contains annotated Reddit posts (in English) of users with depression, anorexia, and control cases. It was created to define early risk detection predictive tasks.

The dataset contains user-level posts, meaning that they correspond to chronologically organized posts of multiple users (a single user can be author of various posts). The dataset is split by use cases: depression (Dataset 1a - depression); and anorexia (Dataset 1b - anorexia). Table 3.2 provides a description of Dataset 1a – depression, and Table 3.3 provides a description of Dataset 1b – anorexia.

Table 3.2. Dataset 1a - depression as described on [93].

Statistics	Train		Test	
	Depressed	**Control**	**Depressed**	**Control**
Num. subjects	135	752	79	741
Num. submissions (posts & comments)	49,557	481,837	40,665	504,523
Avg num. of submissions per subject	367	640.7	514.7	680.9
Avg num. of days from first to last submission	586.43	625.0	786.9	702.5
Avg num. words per submission	27.4	21.8	27.6	23.7

Table 3.3. Dataset 1b - anorexia as described on [93].

Statistics	Train		Test	
	Anorexia	**Control**	**Anorexia**	**Control**
Num. subjects	20	132	41	279
Num. submissions (posts & comments)	7,452	77,514	17,422	151,364
Avg num. of submissions per subject	372.6	587.2	424.9	542.5
Avg num. of days from first to last submission	803.3	641.5	798.9	670.6
Avg num. words per submission	41.2	20.9	35.7	20.9

- *Dataset 2 - Suicidal Ideation* [126]

To obtain this dataset we first elaborated a list of suicide-related sentences. In doing so, we started by collecting a sample of 500 titles of posts published in Reddit's Suicide Watch forum[5]. Posts were mostly written by users with suicidal ideation, so their titles can be considered to be suicide-related sentences.

Sentences were anonymized and then translated to Spanish to avoid the identification of the authors. The sentences were then reviewed by clinic psychologists who added, discarded, and/or adapted the sentences so that they could be used as search terms in Twitter. A subset of 110 phrases were selected.

A total of 98,619 tweets containing the selected sentences were collected for a period of a year, that is, from December 21, 2017, to December 21, 2018. These tweets corresponded to 81,572 Twitter users, with 9,559 users having more than one tweet matched with the search terms. At the same time, for all users, we extracted all their tweets posted within the same search period.

We then followed a two-level annotation process. First, as our intention was to follow a manual labeling process done by clinicians, we selected a random sample of 1,200 users among those who had at least two tweets matching our search phrases. The user names were anonymized, and 3 tags were defined for labeling purposes: 1) control—defining users who on their tweets did not seem to manifest suicidal ideation, users who did not refer to their own conditions, and users who were reporting news or opinions regarding suicide; 2) suicidal ideation risk—labeling users who, judging by their writings, seemed to present suicidal ideation signs; and 3) doubtful—dedicated to cases where psychologists were not sure about labeling them within any of the other categories. A clinician was asked to classify users within these 3 categories based only on the tweets containing the suicide-related keywords. After the labeling process, 73.8% (885/1,200) of users were classified as control cases, 9.6% (115/1,200) were classified as suicidal ideation risk cases, and 16.7% (200/1,200) fell within the

[5] https://www.reddit.com/r/SuicideWatch/

doubtful category. These last doubtful cases were kept to be further considered for evaluating our predictive models.

Then, a second level labeling process for verification was followed for the users tagged as suicidal ideation risk cases. We analyzed more of their profile tweets to confirm their labels. However, annotators noticed that there was a high number of tweets that were not related to suicidal ideation and even sometimes no tweets related to suicide were caught in the sample. To address this issue, we developed a classifier at tweet level to distinguish tweets containing signs of risk from those that were not related at all with suicide. Thus, we could provide the second annotator a summarized version of a user profile, which we call short profile version (SPV), that contained mainly tweets related to suicide and its risk factors.

We built a binary classifier distinguishing 2 classes: 1) suicidal ideation–related tweet and 2) control tweet. To train the model, we chose as instances for the suicide tweet class the tweets of users tagged as suicidal ideation risk cases (513 tweets) and 346 Reddit titles evaluated by the clinicians. For the control tweet class, we selected an equally proportional set of random tweets related to other topics, using Twitter's Sample Tweets application programing interface (API) [159]. A Bag of words model (BoW) was generated. These models represent terms or sequences of terms (n-grams) based on their frequencies on the documents/posts analyzed. We used 1 to 5-grams. Then, we applied principal component analysis (PCA) as a dimensionality reduction method; and logistic regression as a predictive approach with a 10-fold cross-validation procedure. We achieved a model with the following scores: F1 = 0.90, precision (Pr) = 0.91, and recall (R) = 0.89. This is defined as our short profile version classifier (SPVC).

The SPVC was applied to every tweet of the profile of all users labeled as suicidal ideation risk and, for each user, we selected the top 15 suicide-related tweets with the highest predicted probability values given by SPVC. We considered these tweets as the sample to be evaluated by 2 additional annotators: a specialized clinician and a non-specialized annotator. This second annotator was given detailed instructions and information on risk factors related to suicide. The annotators at this stage (second annotation) were asked to classify users into 2 categories: 1) suicidal ideation risk or

2) control, now having more information about each user. We only retained the positive cases (n=84) on which both annotators agreed.

We defined 2 different control groups with the same size as the suicidal ideation risk class:

Focused control group: users writing suicide-related keywords in a non–suicidal ideation risk context, that is, users who trivialize about suicide, news reports, and information regarding the topic; or users who simply manifest their support or opinions to people at risk. Identifying these users is challenging for classification systems but is key in reducing false-positives. These users were chosen at random among the users labeled as control cases during the first annotation process.

Generic control group: a set of Twitter users who might not necessarily use terms related to suicide. These users were selected randomly using the Sample Tweets API [159].

For both control groups, the second annotation process was followed to discard possible cases of users at risk within these samples.

We then obtained a sample of 252 users with a total of 1,214,474 tweets and 305,637 images, from which up to 1,000 images per user were selected for our experiments.

We selected a balanced sample of 84 users presenting signs of suicidal ideation (users at risk), 84 focused control users, and 84 generic control users. Table 3.4 shows the statistics regarding the users belonging to each of the defined groups.

Table 3.4. Dataset 2 - labeled groups' statistics.

Description	Suicidal ideation risk	Focused control	Generic control
Users, n (%)	84 (33.3)	84 (33.3)	84 (33.3)
Tweets collected, n (%)	313,791 (25.8)	766,437 (63.1)	134,246 (11.1)
Tweets per user, median	2,797.5	2,984	716
Terms per tweet, median	11	19	14
Images, n (%)	37,801 (12.4)	251,830 (82.4)	16,006 (5.2)

- *Dataset 3 – Anorexia Nervosa* [124,125]

To build this Spanish dataset we selected keywords and popular hashtags commonly used by ED communities. We also used phrases likely to be used by people undergoing treatment, and terms used by recovered users. These keywords and phrases were manually collected and classified from multiple sources in Spanish and English, including pro-anorexia (proana) blogs, academic publications, and documents made available by the Spanish association against anorexia and bulimia [3,11,49].

In addition, we conducted a survey among volunteers who have recovered from anorexia. The phrases and keywords collected were evaluated and filtered by clinicians that were asked to agree on choosing up to 30 keywords or phrases in Spanish that would lead to reach posts from users with anorexia. Among the terms we find *proana, objective weight (peso objetivo), lose weight (perder peso), body mass index BMI (IMC), sibutramine (sibutramina), my anorexia (mi anorexia),* and *ana and mia (ana y mia)*. We collected 114,627 public tweets from December 21, 2017, to December 21, 2018 containing the search phrases. At the same time, a sample of up to 10,000 tweets from the same search period was collected for each user.

For labeling purposes, we filtered and only considered users with at least three different tweets containing the selected keywords for each category. Among all categories, 645 users met this criterion. Before the submission of the text samples to the annotators, the sample of tweets' texts selected for annotation were anonymized and translated to English to avoid the identification of users based on their writings.

We defined five independent groups of users: 1) AN users that manifest the first stages of the disorder and describe signs and symptoms of AN in their texts, which includes users at both the precontemplation and contemplation stages according to the TTM; 2) a focused control group in which we included users that did not manifest signs of anorexia but use terms related to the disorder in their writings; 3) treatment users that explicitly stated that they have been diagnosed with AN and are in treatment (preparation, action and maintenance stages of the TTM); 4) recovered users who claim they have recovered from AN (termination stage of the TTM); and 5) doubtful cases in

which clinicians were not sure about any of the prior categories. A total of 5 annotators participated in the labeling process: 3 psychologists and 2 psychiatrists. These annotators collaborated closely with organizations specializing in the treatment of EDs.

The final label for a user's set of tweets was assigned if at least three annotators agreed on the assigned label. For cases where an agreement was not met, the users' tweets were categorized as doubtful cases and discarded.

From this first classification approach, a total of 195 users were classified as users with AN, 283 as focused control users, 29 as under treatment users, 18 as recovered users, and 119 as doubtful cases. We performed an inter-annotator agreement analysis and obtained a Light κ coefficient of 0.4751 ($p<.001$), which is the result of the averaged Cohen κ values calculated between each pair of annotators. This approach was chosen over Fleiss κ, as all annotators evaluated every sample. The values obtained suggest a moderate agreement among annotators [100].

As for Dataset 2, in addition to the focused control group, we included another control group consisting of 223 randomly selected users called random control users. These users did not necessarily use terms related to anorexia and were selected also using Twitter's Sample Tweets API.

For our experimental framework, some of the initially labeled users were not further considered in the data set, as they had published less than five tweets during the data collection period, which we considered as not informative enough for our analysis purposes.

A total of 694 users were part of our final data set, which contained data collected from 2,133,110 tweets, including 405,909 images.

Table 3.5 provides relevant information regarding each group in our data collection. For each user, we considered the content from their profiles (tweets) during a year (from December 21, 2017, to December 21, 2018).

To the best of our knowledge, this is the first Spanish data set for the analysis of AN at the user level that considers different stages of the disorder toward recovery.

Table 3.5. Dataset 3 - labeled groups' statistics.

Description	Anorexia nervosa	Treatment	Recovered	Focused control	Random control
Users: n (%)	171 (24.6)	27 (3.9)	18 (2.6)	271 (39)	207 (29.8)
Tweets collected: n (%)	434,615 (20.4)	8,317 (0.4)	52,578 (2.5)	1,109,861 (52)	447,739 (21)
tweets per user: median	1,239	1,748	2,036.5	2,608	873
Terms per tweet: median	14.00	14.00	13.50	12.50	19.00
Images: n (%)	40,142 (9.9)	6,584 (1.6)	4,202 (1)	298,488 (73.5)	56,493 (13.9)

- *Dataset 4 – all use cases* [127]

This is a Reddit dataset (in English) that was created to address a multi-class predictive task. Data was collected from a group of selected subreddits and it was automatically labeled. We considered subreddits addressing suicidal ideation (*Suicidewatch*), depression (*depression*), alcoholism (*alcoholism*), and eating disorders (*eating_disorders, bulimia,* and *EatingDisorders*).

As it was our intention to consider only posts of users living with the selected conditions and not control cases within the subreddits, we applied an automatic labeling approach where a post was first assigned the label of the subreddit it belonged to.

Later, a first filtering approach was applied such that only posts with self-references were considered. With this purpose, we only kept posts containing keywords and phrases such as: my alcoholism, I was diagnosed, I′m anorexic, etc. From the starting 282,448 posts, with this filtering approach we kept only 13,174 posts.

Taking into account the characteristics of a multiclass task, we proceeded to discard posts of users with possible comorbidities. For the posts belonging to a given class, we did not keep posts with main general terms that describe other classes (e.g., for the alcoholism group we discarded posts containing the main terms: *depression, anorexia, bulimia, eating disorders* and *suicide*). We considered 9 main terms in total for this step.

After the filtering process, only 11,124 posts were kept. Finally, the keywords used for the first filtering approach were

removed from 70% of the posts so that the keywords used during the data gathering would not interfere in the predictive models' behavior.

To collect control posts (CON), we took into account posts from subreddits where all types of posts were published. We considered posts from 18 randomly selected subreddits such as: *sports, celebs, books, fan theories, space, science, medical school, travel, history, economics, ask engineers, art fundamentals, lectures, unsolved mysteries, tales from call centers, law, legal advice* and *shower thoughts*. To discard posts that could be related to any of the issues studied, we deleted those containing self-references related to the mental conditions addressed. A total of 20,057 control posts were considered for our experimental approach.

Given that we address binary and a multiclass tasks with this dataset, we define 2 sub-datasets: Dataset 4a - multiple includes the labeled posts that correspond to the depression (DEP), eating disorders (ED), suicidal ideation (SUI) and alcoholism (ALC) classes. Dataset 4b – mental contains the posts that correspond to Control (CON) cases, and those of what we define as the MEN class, which consists on the union of all the groups (DEP + ED + SUI + ALC) of Dataset 4a – multiple, meaning it is a superset of it. Table 3.6 describes the classes and statistics of both datasets.

Table 3.6. Dataset 4 - labeled groups' statistics.

Dataset	Class	Posts, n (%)	Terms per post, median
4a-multiple	Suicide (SUI)	7,075 (63.6)	136
	Depression (DEP)	3,015 (27.1)	177
	Alcoholism (ALC)	250 (2.3)	241
	Eating disorders (ED)	784 (7.0)	191
4b-mental	Mental Conditions (MEN)	11,124 (35.7)	152
	Control (CON)	20,057 (64.3)	141

3.4.2 Evaluation

To assess the framework proposed we develop characterization, predictive and recommendation models for multiple use cases. The contributions for each module are evaluated using different measures depending on the methods developed.

Characterization approaches use mostly methods to compare the values of features between the groups defined in each dataset. Thus, we take into account p-values as measures to assess that the differences observed did not occur just by random chance [125–127].

Regarding predictive models, we evaluate model's performances using metrics such as Precision, Recall, F1-Score, and AUC [98,125–129]. We also measure the delay in emitting a decision on early risk detection approaches [98,129] with the ERDE measure [91].

As we propose new text representations (enhanced word embeddings) for the detection of mental disorders, we also make use of similarity measures (cosine similarity) and visual evaluation approaches (bi-dimensional vector plots) [127,128].

Finally, to assess our contact recommendation module we measure Precision (P), Recall (R) and the Mean Average Precision (MAP) at a given number (k) of ranked users recommended.

The recommendation proposal is evaluated both in the context of a social platform taking into account data collected from users; and it is also evaluated with the participation of volunteers in treatment, in a simulation of a OSN context.

We also propose a new evaluation measure that combines MAP and the ratio of harmless users suggested that the target user is actually likely to follow.

3.5 Ethical assessment

The analysis of data provided by social networks to detect health problems and assist clinicians is an open issue, not uncontroversial. The aim of our proposal, however, is to shed light on the real capabilities of these systems in specific practical applications. Before such systems become available, a careful risk-benefit assessment along with a proper analysis of applicable legal framework compliance and the potential

threats to users' privacy and civil liberties shall be conducted [25,74,173].

The work reported in this thesis has been approved by the ethical review board of Pompeu Fabra University (CIREP Approval number: 162). Most of the characterization and predictive tools' evaluation approaches have been part of observational studies.

Also, the evaluation of the recommendation approach proposed has involved the participation of volunteers that agreed to participate in our evaluation tasks. Confidentiality and participation agreements have been signed by all the annotators and volunteers taking part of our studies.

To avoid processing and storing personal or sensitive data, a proper process of data transformation and anonymization was followed. We only stored the extracted transformed features.

In terms of reproducibility, the policies on the distribution of the data collected through the social platforms' APIs is respected. No information that could lead to the identification of the users included in our study will be shared, as we did not store any personal information. However, the values of the features calculated are available upon reasonable request and after a proper evaluation of the use purpose.

Chapter 4

CHARACTERIZATION OF MENTAL HEALTH STATES

4.1 Introduction

In this section we report the details of our work dedicated to mental health states characterization. We first explore the main elements that characterize mental health issues in general on social platforms, and define a comparative analysis of the elements that characterize each of our use cases. Then we study in depth two use cases selected: anorexia nervosa and suicidal ideation. Finally, we discuss our findings and their relevance for the development of tools that can assist users with mental health issues through social platforms.

4.2 Mental health states characterization

4.2.1 Introduction

We first report our findings regarding exploratory work performed over Reddit data for the characterization of eating disorders (ED), suicidal ideation (SUI), depression (DEP) and alcoholism (ALC) [127]. This approach makes use of Dataset 4 (See Table 3.6) which is constituted by two sub-datasets, one dedicated to identify features that characterize specific conditions, and that distinguish a condition from another (Dataset 4a – Multiple); and another that is used to explore

elements that are common to multiple mental disorders and that distinguish them from control cases (Dataset 4b – Mental).

We perform a comparative analysis of the dataset posts to characterize the mental conditions studied using lexicons dedicated to five themes: affective processes and emotions, personal concerns and biological processes, linguistic elements, vocabulary related to risk factors, and topics of interest. We assume that the different groups (depression (DEP), eating disorders (ED), suicidal ideation (SUI) and alcoholism (ALC)) do not tackle the same topics and that they do not use the same vocabulary. We use statistical tests to check this hypothesis.

Prior work has been dedicated to the detection of mental disorders on social media [19,39,63,75,90,135,146]. Most of it has been focused on the analysis of a single condition, which is usually compared to control cases [39,63,75,90]. Other studies have considered different risk levels over a single condition [146]; whereas only a few publications have been dedicated to the detection [19] and comparative analysis of multiple mental conditions, which are likely to be characterized by similar signs and symptoms [135]. Through our work, we do a further exploration of the linguistic dimensions, affective processes and emotions, personal concerns, vocabulary related to risk factors [126], and topics of interest linked to each condition, and define a method to identify the terms or n-grams that are highly related to them.

4.2.2 Comparative analysis

To identify the elements that characterize and differentiate each of the conditions considered, we perform a comparative analysis of the types of posts studied. With this purpose, we consider different psychological and linguistic perspectives that correspond to lexicons' categories, where each category is composed by a set of terms.

We generate numeric features for each of the categories analyzed within each perspective. To do so, for a given post we counted the frequency of terms belonging to each of the categories of the dictionaries, then the frequency was normalized by the size (in number of terms) of the full post. This approach was followed for all the lexicons' perspectives.

For the comparison of the groups analyzed (DEP, ED, SUI, and ALC),we apply non-parametric tests after verifying that our features do not follow a normal distribution and that there is no homogeneity of variance for most of them. We first verified that there were features with significant differences among all the groups using Kruskal-Wallis' test [83].

Once we found there were features with significant differences, we performed Mann-Whitney U's test [96] to check if the difference is significant for those features between pairs of groups. We also use this test to compare mental conditions (MEN) and control (CON) cases.

We analyzed 5 different perspectives as defined in the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary [23], which categorizes words in psychologically meaningful perspectives. We also consider other domain specific perspectives. The description of these perspectives and the results obtained for our comparative analysis are as follows.

a) Affective processes and emotions:

To address these perspectives we consider some of the LIWC's lexicon categories in addition to the categories described in EmoLex [104], which is a dictionary that associates terms to negative and positive sentiments, along with eight basic emotions: anger, anticipation, trust, fear, surprise, sadness, disgust and joy.

A total of 23 categories are analyzed in this perspective. Table 4.1 shows the mean score values computed for selected categories from [127] over the set of writings of each of the groups compared (MEN, CON, SUI, DEP, ED, ALC) and the p-values with the level of significance for each pair of classes compared using Mann-Whitney U's test.

The averaged values per group are reported since the median values are zero for a great number of features, as there are many categories with terms that can be found on very few writings.

Results show that there are features with high significant differences between the pairs of groups. As expected, this is notably true between the mental conditions (MEN) and control (CON) groups, where the differences were significant for all the 23 categories, confirming the quality of Dataset 4b - mental.

Table 4.1. Comparative results (means and p-values) between groups according to the affective processes and emotions perspective.

Categories	Mean values per group						p-values and significance level (Mann-Whitney U)						
	MEN	CON	SUI	DEP	ED	ALC	MEN - CON	SUI -DEP	SUI - ED	SUI - ALC	DEP - ED	DEP - ALC	ED - ALC
Fear	3.30E-02	1.56E-02	3.67E-02	2.60E-02	2.78E-02	2.31E-02	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.002 **	<0.001 ***
Disgust	17.6E-03	7.74E-03	18.9E-03	15.0E-03	18.1E-03	15.2E-03	<0.001 ***	<0.001 ***	0.249	0.002 **	<0.001 ***	0.422	<0.001 ***
Joy	1.65E-02	1.46E-02	1.61E-02	1.69E-02	2.05E-02	1.62E-02	<0.001 ***	<0.001 ***	<0.001 ***	0.052	<0.001 ***	0.391	0.003 **
Anger	2.23E-02	1.21E-02	2.36E-02	2.11E-02	1.42E-02	1.79E-02	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.001 **	<0.001 ***
Surprise	10.1E-03	8.67E-03	10.2E-03	9.83E-03	12.2E-03	8.49E-03	<0.001 ***	0.005 **	<0.001 ***	0.46	<0.001 ***	0.165	<0.001 ***
Sadness	3.71E-02	1.32E-02	3.92E-02	3.60E-02	2.55E-02	2.65E-02	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.395
Positive emotions	3.64E-02	4.64E-02	3.51E-02	3.78E-02	4.50E-02	3.55E-02	<0.001 ***	<0.001 ***	<0.001 ***	0.24	<0.001 ***	0.063	<0.001 ***
Negative emotions	5.24E-02	2.64E-02	5.43E-02	5.01E-02	4.35E-02	4.65E-02	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.030 *	0.089

(***p<0.001, **p<.01, *p<.05)

Emotions such as anger, fear, disgust and sadness are expressed more on texts of users with suicidal ideation in comparison to texts of users with depression. These users (suicidal ideation) are the ones that express more negative emotions.

Users with eating disorders are the most positive ones compared to the other groups within the conditions analyzed, and this can be because eating disorders such as anorexia and bulimia can be characterized by the Transtheoretical Model of Health Behavior Change, where people at the pre-contemplation stage are enthusiastic about their weight loss, and the social support they receive.

The groups having the least differences between each other are the depression and alcoholism groups as shown by the non-significant results of the tests for several categories (in Table 4.1 we show the results for categories such as disgust, joy, surprise, etc.). In Figure 4.1 (left), which presents a comparison of the emotions (Emolex) scores according to Plutchik's wheel [119], we can notice that sadness and fear are the emotions that mostly characterize users with mental conditions compared to the control group.

b) Personal concerns and biological processes:

Using LIWC, we also explore lexicon categories that are related to daily activities and concerns of users through general terms related to religion, work, leisure, money, health, and biological processes.

A total of 12 categories were analyzed. Table 4.2 reports the comparative results obtained for these categories. Control writings obtain the highest scores for the categories work, money, and home and the lowest for biological processes, body, or health.

We can notice a high mean value for the usage of terms related to death and sexuality for the suicide class and, for the categories: body, ingest and biological processes in the eating disorders class. These last two categories also characterize the alcoholism class, which obtains the highest mean scores for both of them with the leisure category. Also, the ED class obtains the highest score for the achievement category and the lowest score for the religion and death categories in comparison to the other conditions.

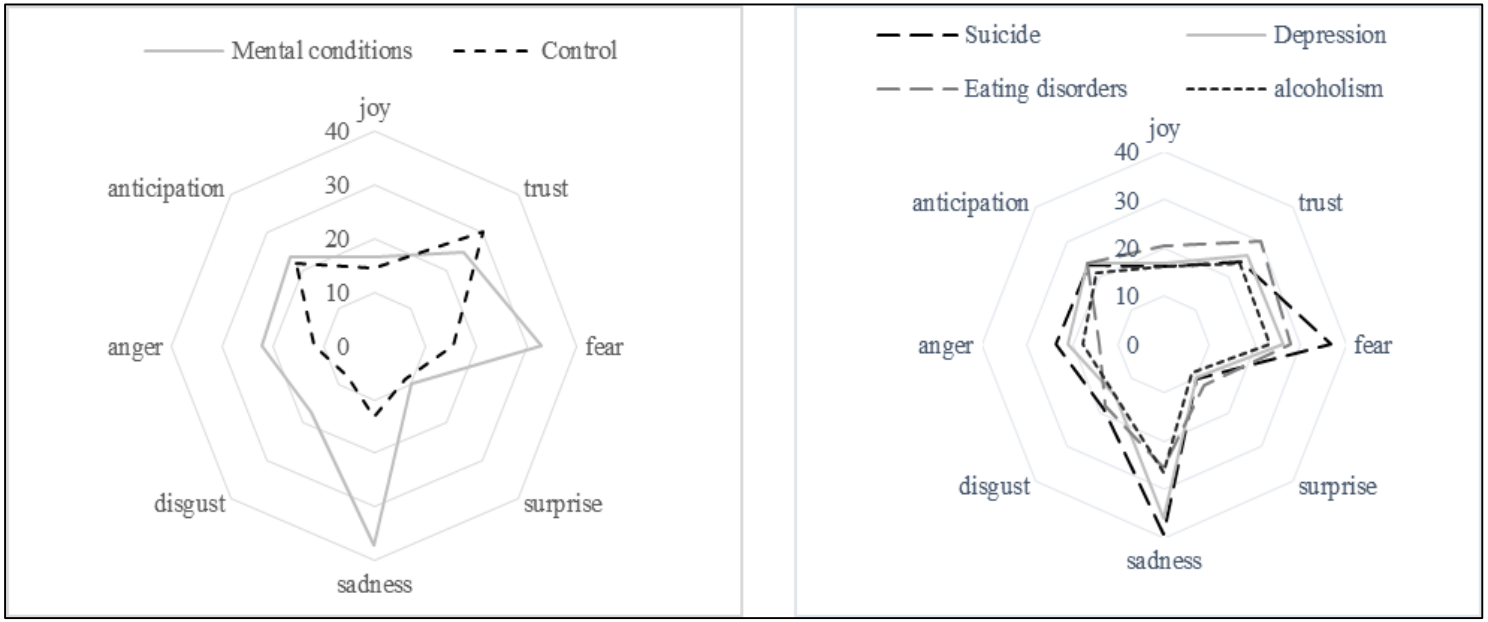

Figure 4.1. Emotions (Emolex) scores according to the basic emotions of Plutchik's wheel. The scores from Table 4.1 were multiplied by 1000 to ease the visualization.

Table 4.2. Comparative results (means and p-values) between groups according to the personal concerns and biological processes perspective.

Categories	Mean values per group						p-values and significance level (Mann-Whitney U)						
	MEN	CON	SUI	DEP	ED	ALC	MEN – CON	SUI – DEP	SUI – ED	SUI – ALC	DEP– ED	DEP– ALC	ED – ALC
Work	3.87E-02	10.8E-02	3.33E-02	5.03E-02	3.41E-02	4.12E-02	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.437	<0.001 ***
Achievement	3.72E-02	4.32E-02	3.47E-02	3.72E-02	5.47E-02	3.96E-02	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.003 **	<0.001 ***
Leisure	1.79E-02	3.46E-02	1.43E-02	1.81E-02	1.70E-02	9.70E-02	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.425	<0.001 ***	<0.001 ***
Home	8.24E-03	14.1E-03	7.71E-03	9.46E-03	7.76E-03	8.40E-03	<0.001 ***	<0.001 ***	0.299	<0.001 ***	0.001 **	0.169	0.004 **
Money	8.33E-03	40.8E-03	8.61E-03	7.47E-03	5.84E-03	13.6E-03	<0.001 ***	0.007 **	0.408	<0.001 ***	0.053	0.003 **	<0.001 ***
Religion	3.61E-03	4.54E-03	3.61E-03	3.40E-03	1.58E-03	5.90E-03	<0.001 ***	0.466	<0.001 ***	0.008 **	<0.001 ***	0.011 *	<0.001 ***
Sexual	14.1E-03	4.19E-03	18.4E-03	8.12E-03	4.28E-03	4.83E-03	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.147	0.006 **
Death	33.9E-03	5.84E-03	50.1E-03	7.02E-03	1.34E-03	3.18E-03	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***
Biological processes	8.16E-02	3.04E-02	7.20E-02	7.61E-02	15.3E-02	18.7E-02	<0.001 ***	0.003 **	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***
Body	17.9E-03	8.37E-03	18.5E-03	16.7E-03	20.6E-03	14.6E-03	<0.001 ***	0.158	<0.001 ***	0.166	0.005 **	0.243	0.169
Ingest	13.4E-03	4.61E-03	4.37E-03	6.13E-03	85.6E-03	122E-03	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***
Health	4.09E-02	1.42E-02	3.24E-02	4.60E-02	7.69E-02	8.40E-02	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.124

(***p<.001, **p<.01, *p<.05)

Concerning the SUI group, it obtains the lowest scores in the work, achievement, and leisure categories whereas the depression class has the second highest value for the usage of terms related to death, and this score is significantly higher in comparison to the eating disorders and alcoholism classes as confirmed by the p-value.

c) Linguistic elements:

This perspective addresses the usage of grammatical and syntactical elements such as verbs, adverbs, pronouns, articles, and prepositions. It also considers the different verbal times and pronoun types. We use LIWC for this perspective as well.

We consider 16 features of this type. In Table 4.3, we can observe a selection of these features. Writings of users of the MEN group, in comparison to the CON group, tend to have more first-person singular pronouns, use more negations, adverbs, verbs, and past and present verb tenses. In comparison to the other conditions, the suicide group is characterized mainly by the usage of pronouns, especially first-person singular pronouns. It is also characterized by the reduced usage of second person pronouns, past verb tenses and articles; and the high usage of negations, and present and future verb tenses. A characteristic of the depression class is the high usage of third person plural pronouns in comparison to the other conditions. It also gets scores significantly lower than the suicide class but also significantly higher than the ED and ALC classes in the following categories: verbs, personal pronouns, and present verb tense. The ED group is characterized by the usage of first-person plural pronouns which is significantly higher than the SUI class but significantly lower than the DEP and ALC classes. Finally, the ALC group is characterized by a low usage of adverbs, and a high usage of articles and prepositions.

d) Domain related vocabulary:

We study lexical categories related to eating disorders, self-loathing, self-injuries, explicit suicidal ideation references, substance abuse, lack of social support, and discrimination or abuse. These categories were taken from the outcome of our work dedicated to the analysis of suicidal ideation [126] (described in Section 4.3).

Table 4.3. Comparative results (means and p-values) between groups according to the linguistic elements perspective.

Categories	Mean values per group						p-values and significance level (Mann-Whitney U)						
	MEN	CON	SUI	DEP	ED	ALC	MEN-CON	SUI - DEP	SUI - ED	SUI - ALC	DEP- ED	DEP-ALC	ED - ALC
First person singular pronouns	13.4E-02	5.18E-02	14.3E-02	12.0E-02	11.7E-02	11.4E-02	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.091	0.012 *	0.085
First person plural pronouns	1.70E-03	6.71E-03	1.47E-03	2.29E-03	1.60E-03	2.54E-03	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.017 *	<0.001 ***
Second person pronouns	4.68E-03	11.0E-03	4.08E-03	5.21E-03	4.81E-03	5.21E-03	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.303	0.022	0.017 *
Third person plural pronouns	7.61E-03	13.7E-03	7.56E-03	8.23E-03	5.92E-03	4.81E-03	<0.001 ***	<0.001 ***	0.027 *	0.046 *	<0.001 ***	0.001 **	0.315
Negations	1.86E-02	1.03E-02	2.09E-02	1.60E-02	1.35E-02	1.28E-02	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.002 **	0.239
Adverbs	6.57E-02	4.69E-02	6.65E-02	6.64E-02	6.64E-02	5.96E-02	<0.001 ***	0.108	0.136	<0.001 ***	0.337	<0.001 ***	<0.001 ***
Articles	3.79E-02	6.79E-02	3.68E-02	3.87E-02	4.03E-02	5.11E-02	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.030 *	<0.001 ***	<0.001 ***
Verbs	1.52E-01	1.27E-01	1.56E-01	1.50E-01	1.37E-01	1.35E-01	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.346
Personal pronouns	15.6E-02	9.98E-02	16.3E-02	14.5E-02	13.8E-02	13.5E-02	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.15
Prepositions	1.22E-01	1.28E-01	1.19E-01	1.23E-01	1.25E-01	1.36E-01	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.107	<0.001 ***	<0.001 ***
Past verb tense	3.38E-02	3.22E-02	3.22E-02	3.78E-02	3.87E-02	4.04E-02	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.114	0.028 *	0.135
Present verb tense	9.67E-02	7.46E-02	10.0E-02	9.35E-02	8.32E-02	7.80E-02	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.046 *
Future verb tense	9.77E-03	9.09E-03	11.7E-03	6.60E-03	6.05E-03	6.58E-03	0.366	<0.001 ***	<0.001 ***	<0.001 ***	0.3	0.085	0.05

(***p<.001, **p<.01, *p<.05)

We also considered the categories defined by Arseniev *et al.* [11] with terms related to anorexia nervosa and its symptoms. These categories are: anorexia promotion, body image, body weight, caloric restrictions, compensatory behaviors, and exercise. We also consider names of antidepressants.

Results regarding this perspective are shown in Table 4.4, with some selected categories among the 23 studied. We can notice that there are significant differences for all categories between the MEN and CON groups, with higher mean values for the former. As expected, when compared to the other mental conditions' groups, the SUI group obtains a very significantly high score for the explicit suicide category; it also obtains the lowest mean value for the food and meals category and, the second lowest score for the explicit depression category with highly significant differences with the remaining classes. The categories that characterize the DEP group are the explicit depression and antidepressants, while for the ED group are those related with food and meals, caloric restriction, anorexia promotion, eat verb, body image, binge eating, body weight, compensatory behavior and laxatives. Regarding the ALC group, one can notice a high value for the substance abuse category, as expected, but also the lowest mean value for the hate category when compared with the other conditions.

e) Topics of interest:

Using Empath [52], which generates and validates lexical categories using a corpus with 1.8 billion words, we retain 200 prebuilt topics such as sports, social media, music, and politics, among others. Figure 4.2 shows only the top 20 Empath topics (categories) having the most significantly different values ($p<.05$) between each pair of classes compared, including the mental conditions and control classes. The mean value for each class compared for each topic is shown.

We can observe that the SUI group, compared to the other conditions' groups is characterized by addressing topics such as: kill, crime, prison, weapon, war, fight, aggression, negative emotions, and hate. The ED group is characterized by topics such as food, eating, cooking, restaurant, shopping, and strength, which can easily be linked to the condition.

Table 4.4. Comparative results (means and p-values) between groups according to the domain related vocabulary perspective.

Categories	Mean values per group						p-values and significance level (Mann-Whitney U)						
	MEN	CON	SUI	DEP	ED	ALC	MEN - CON	SUI - DEP	SUI - ED	SUI - ALC	DEP - ED	DEP - ALC	ED - ALC
Explicit suicide	29.3E-04	1.34E-04	43.5E-04	7.18E-04	1.39E-04	7.39E-04	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.017 *	0.043 *
Food and meals	15.7E-04	5.52E-04	4.08E-04	9.49E-04	142E-04	10.2E-04	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.059	<0.001 ***
Caloric restriction	6.10E-04	1.10E-04	1.75E-04	3.81E-04	51.4E-04	12.1E-04	<0.001 ***	<0.001 ***	<0.001 ***	0.087	<0.001 ***	0.448	<0.001 ***
Anorexia promotion	89.5E-05	3.94E-05	14.9E-05	34.8E-05	944E-05	13.1E-05	<0.001 ***	<0.001 ***	<0.001 ***	0.264	<0.001 ***	0.058	<0.001 ***
Eat verb	62.7E-05	5.08E-05	9.80E-05	23.1E-05	676E-05	12.3E-05	<0.001 ***	<0.001 ***	<0.001 ***	0.044 *	<0.001 ***	0.32	<0.001 ***
Explicit depression	156E-05	3.47E-07	21.7E-05	522E-05	85.0E-05	18.5E-05	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.008 **
Hate	18.3E-04	1.47E-04	20.4E-04	13.1E-04	11.7E-04	2.31E-04	<0.001 ***	0.017 *	0.032 *	<0.001 ***	0.263	<0.001 ***	<0.001 ***
Substance abuse	52.7E-04	7.74E-04	27.2E-04	27.7E-04	24.8E-04	114E-03	<0.001 ***	0.020 *	0.341	<0.001 ***	0.073	<0.001 ***	<0.001 ***
Body image	23.7E-05	2.48E-05	14.2E-05	3.10E-05	223E-05	1.60E-05	<0.001 ***	0.083	<0.001 ***	0.231	<0.001 ***	0.367	<0.001 ***
Binge eating	72.1E-05	3.12E-05	5.60E-05	7.00E-05	822E-05	153E-05	<0.001 ***	0.172	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***
Body weight	46.8E-05	12.0E-05	9.90E-05	26.4E-05	468E-05	19.7E-05	<0.001 ***	<0.001 ***	<0.001 ***	0.002 **	<0.001 ***	0.448	<0.001 ***
Antidepressants	770E-06	5.25E-06	268E-06	200E-05	363E-06	0.000	<0.001 ***	<0.001 ***	0.459	0.041 *	<0.001 ***	<0.001 ***	0.044 *
Compensatory behavior and laxatives	31.3E-05	1.08E-05	13.1E-05	4.30E-05	355E-05	9.40E-05	<0.001 ***	0.062	<0.001 ***	0.317	<0.001 ***	0.12	<0.001 ***

(***p<.001, **p<.01, *p<.05)

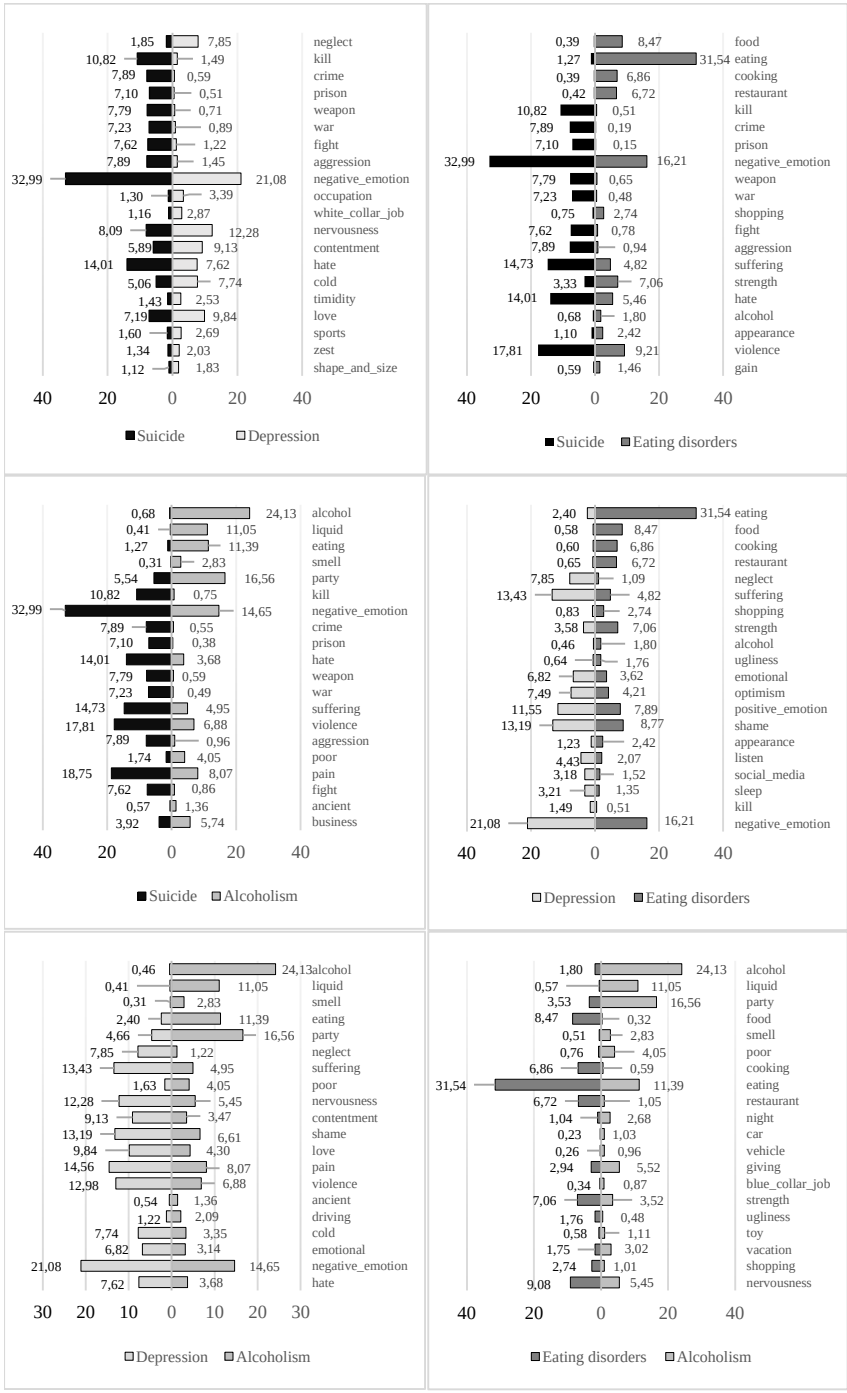

Figure 4.2. Top 20 Empath topics with most significantly different values ($p<.05$) between each pair of classes compared (multi-class task). The mean value for each class compared and topic is shown.

The topics that characterize the ALC group are alcohol, liquid, party, smell, and poor. This last one is a topic that normally implies the usage of terms related to economic issues, but in this case the topic is likely to be representative because within its terms the word *alcoholism* can be found. The DEP group is characterized only by the neglect topic compared to all the other conditions, this topic considers terms such as: *depressed, loneliness, fear, depression, loathing, hopelessness* and *suffering*.

When the DEP group is compared to the ED and ALC groups, we can observe that suffering, emotional, shame and negative emotion are topics that obtain significantly higher scores for the DEP group. Regarding the SUI vs. DEP class, we can see that the depression group expresses more feelings of contentment, love and zest, and it also addresses more topics related to daily activities such as white-collar jobs, occupations (professions), and sports. Notice too that when the ED and ALC groups are compared, the ALC group addresses more leisure related topics such as party, night, car, and vacation. Finally, compared to the control group (Figure 4.3), the mental conditions group obtains higher mean values on topics that address feelings and emotions.

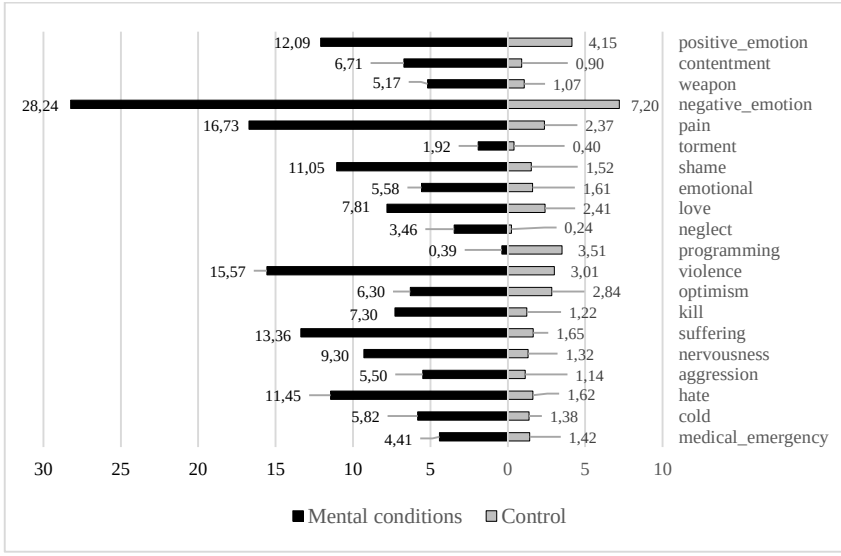

Figure 4.3. Top 20 Empath topics with most significantly different values (p<.05) between the classes compared (binary task). The mean value for each class and topic is shown.

These findings confirm our hypothesis according to which the vocabulary used by the different groups is not the same. Specific topics, with their corresponding terms, are highly addressed by a given group, such as caloric restriction by the eating disorders' group, they also reveal less obvious associations, which suggest that such terms could be efficiently exploited as predictive features to automatically determine the belonging of a user to a group in function of their writings.

4.2.3 Detection of relevant n-grams

In this section we propose algorithms for the detection of unigrams and bigrams that are relevant for a given group compared to other groups. For instance, we can identify the n-grams that characterize the SUI group when compared to the DEP, ED and ALC groups. These terms are later used to improve text representations (word embeddings) used for predictive tasks (Section 5.3.3), as the approach can be used to define n-grams or terms that are predictive for a given class (group).

This is why we denote the output of the algorithm to be lists of terms for each given class. We address two cases, one where there are only two groups to compare, and the second where there are more than two groups to address.

We first define an algorithm that addresses multiple groups. Each group is denoted as a class, and each social media post is labeled as a document that corresponds to a given class. An instance for our use cases is a post from the dataset that has been labeled to correspond to the SUI class. In this context we define Algorithm 4.1.

We summarize the process in 4 main steps: *first*, based on X^2 [94], we aim to identify the classes for which each unique {1-2}-gram of the corpus is more relevant. Given the labeled documents as input, for the X^2 definition of relevant terms for each class, a BoW model with a Boolean representation denoting the existence of a term in a given document is generated (Boolean_matrix), along with the classes (labels) to which each document belongs. Then, we proceed to calculate the X^2 scores for each term and class. As for a given term t, a X^2 score is obtained for each class c_n in the list of existing classes C and stored ($X^2scores_t$), choosing the class c_n for

Algorithm 4.1. Relevant n-grams detection for multiple groups/classes compared

Input: labeled_documents, Classes list C
Output: positive predictive terms lists $c_n\_predictive$

1. $Boolean\_matrix \leftarrow$ generate_Boolean_matrix($labeled\_documents$)
2. $X^2\_scores \leftarrow$ calculate_ X^2_scores($Boolean\_matrix$)
3. for every term t in X^2_scores
4. $max\_score \leftarrow \max(X^2\_scores_t)$
5. $c_n \leftarrow$ class to which max_score corresponds
6. append t to rel_c_n
7. end for
8. $Tf.Idf\_model \leftarrow$ generate_TF − IDF_representation($labeled\_documents$)
9. for every class c_n in C
10. for every term t in rel_c_n
11. $is\_relevant \leftarrow$ True
12. for every class c_m in $C\backslash\{c_n\}$
13. $Pval \leftarrow$ P_value_Mann_Whitney_U ($Tf.Id\_model_{t,c_n}, Tf.Id\_model_{t,c_m}$)
14. if $Pval < 0.001$
15. $mean_{t,c_n} \leftarrow \text{mean}(Tf.Id\_model_{t,c_n})$
16. $mean_{t,c_m} \leftarrow \text{mean}(Tf.Id\_model_{t,c_m})$
17. if $mean_{t,c_n} < mean_{t,c_m}$
18. $is\_relevant \leftarrow$ False
19. break
20. end if
21. else
22. $is\_relevant \leftarrow$ False
23. break
24. end if
25. end for
26. if $is\_relevant$
27. add t to $c_n\_predictive$
28. end if
29. end for
30. end for

which *t* obtains the highest X^2 score (max among the scores for each class in $X^2scores_t$) and we add *t* to the list of relevant terms of c_n (rel_c_n) according to the X^2 test (Steps 1 to 7 in Algorithm 4.1). By this way, a list of relevant terms is generated for each of the classes in *C,* and each term is relevant for one single class.

As every {1-2}-gram of the vocabulary is defined as relevant for a given class regardless of having a very low X^2 score, it is important to select only the most relevant terms for a class, *i.e.,* a subset of all the terms relevant to c_n. Therefore in the *second* step we proceed to create a TF-IDF representation of the posts (documents) for all the terms {1-2}-grams of the corpus. A TF-

IDF model provides a weight for a term in a document. For all the thesis' experiments involving such models, for the TF we use the log normalization, and for the IDF we use the inverse frequency as described in [15]. Then, for each class c_n in C we apply Mann Whitney U's test [96] for each term t belonging to rel_c_n in order to compare the TF-IDF scores for t of all the documents that correspond to c_n and the TF-IDF scores for t of the documents belonging to each one of the remaining classes in C ($C\backslash\{c_n\}$). This step corresponds to the statements 8 to 13 in Algorithm 4.1.

In the *third* main step, for those pairs of classes where the P-value obtained for a term t by the Mann Whitney U's test is lower than a given threshold (0.001 in our use case), we calculate the mean TF-IDF score obtained by t for each class of the pair, and then we pick the class for which the mean TF-IDF value is the highest as the one for which t is relevant (steps 13 to 26 in Algorithm 4.1).

At the *fourth* main step, if t is relevant for the same class c_n on all its comparisons with the remaining classes in C, then it is kept and added to the list of positive predictive terms for this class ($c_n\_predictive$) (Steps 27 to 29 in Algorithm 4.1).

Table 4.5 shows the list of the top 15 most relevant terms for each class among all the relevant terms selected after applying the approach described in Algorithm 4.1 to Dataset 4a - multiple. For the alcoholism class, we observed that despite having a reduced number of posts it is a class that can be characterized by a large number of terms, whereas the suicide class, despite having the largest number of writings, does not have a large amount of unique distinguishable terms.

When addressing only two classes we propose a variation of the prior approach but we consider that for two groups, the X^2 resulting predictive terms are the same for both classes. The steps to obtain the predictive terms for this task type are shown in Algorithm 4.2.

We first consider the same initial main step as for the case of multiple classes except that for this binary case, we define an X^2 score threshold based on the distribution of the scores of all the terms in order to keep only relevant terms. Then, these terms, regardless of the class they are relevant for (as it is not known through the X^2 test) are added to a list of binary relevant terms (*binary_rel_terms*) (steps 1-3 of Algorithm 4.2).

Table 4.5. List of the top 15 most relevant terms when comparing the SUI, DEP, ED and ALC groups (multiple groups' case).

Class	SUI	DEP	ED	ALC
Terms' number	11	6	45	56
Terms	*Kill* *Suicide*	*Depression* *Anxiety*	*Eating* *Eating disorder*	*Alcoholism* *Alcohol*
	Die	*Depressed*	*Bulimia*	*Alcoholic*
	Want die	*Depression anxiety*	*Purging*	*Drinking*
	Killing	*Energy*	*Ed*	*Drink*
	Live	*Mental health*	*Purge*	*Sober*
	Dead	*Sad*	*Weight*	*AA*
	Anymore	-	*Recovery*	*Beer*
	Just want	-	*Food*	*Sobriety*
	Cares	-	*Anorexia*	*Drank*
	Care	-	*Eat*	*Liquor*
	Kill myself	-	*Binge*	*Drinks*
	-	-	*Calories*	*Drunk*
	-	-	*Bulimic*	*Stop drinking*
	-	-	*Binging*	*Beers*
	-	-	*Restricting*	*Drinking problem*

Algorithm 4.2. Predictive terms' lists generation for binary classification tasks

Input: $labeled\_documents$
Output: predictive terms lists $c_n\_predictive$

1. $Boolean\_matrix \leftarrow$ generate_Boolean_matrix($labeled\_documents$)
2. $X^2\_scores \leftarrow$ calculate_ X^2_scores($Boolean\_matrix$)
3. $binary\_rel\_terms \leftarrow$ terms that obtain X^2scores over a threshold
4. $Tf.Idf\_model \leftarrow$ generate_TF − IDF_representation($labeled\_documents$)
5. for every term t in b$inary_rel_terms$
6. $Pval \leftarrow$ P_value_Mann_Whitney_U($Tf.Id\_model_{t,c_1}, Tf.Id\_model_{t,c_2}$)
7. if $Pval < 0.001$
8. $mean_{t,c_1} \leftarrow$ mean($Tf.Id\_model_{t,c_1}$)
9. $mean_{t,c_2} \leftarrow$ mean($Tf.Id\_model_{t,c_2}$)
10. if $mean_{t,c_1} > mean_{t,c_2}$
11. add t to $c_1\_predictive$
12. else
13. add t to $c_2\_predictive$
14. end if
15. end if
16. end for

Later, to identify the class for which the terms in the *binary_rel_terms* list are predictive or relevant, and to discard

terms that are not relevant enough, we execute the main steps 2 to 4 of the approach for the multi-class task.

When addressing only two classes we propose a variation of the prior approach but we consider that for two groups, the X^2 resulting predictive terms are the same for both classes. The steps to obtain the predictive terms for this task type are shown in Algorithm 4.2.

We first consider the same initial main step as for the case of multiple classes except that for this binary case, we define an X^2 score threshold based on the distribution of the scores of all the terms in order to keep only relevant terms.

Then, these terms, regardless of the class they are relevant for (as it is not known through the X^2 test) are added to a list of binary relevant terms (binary_rel_terms) (steps 1 to 3 of Algorithm 4.2).

Later, to identify the class for which the terms in the binary_rel_terms list are predictive or relevant, and to discard terms that are not relevant enough, we execute the main steps 2 to 4 of the approach for the multi-class task.

We consider for this case that for the second main step, Mann Whitney U's test is applied for each term in the *binary_rel_terms* list and that the comparison is done between the TF-IDF scores of the documents according to the respective class they belong to (steps 4 to 6 of Algorithm 4.2). In this sense, if the p-value threshold is met for a given term, then it is directly added to the list of predictive terms of the class for which it obtains the greatest mean TF-IDF score c_n_predictive (steps 5 to 13 in Algorithm 4.2).

Table 4.6 shows the top 15 most predictive terms obtained after applying Algorithm 4.2 to the case where only the MEN and CON classes are considered (Dataset 4b – mental).

4.3 Characterization of suicidal ideation

4.3.1 Introduction

In this section we describe our work dedicated to the characterization of suicidal ideation on social platforms. The work described in this section corresponds to our findings reported in [126]. For this case we introduce the analysis of images in addition to text features.

Table 4.6. List of the most relevant/predictive terms for each class when comparing only the MEN and CON groups.

MEN	CON
feel	*company*
life	*customer*
kill	*calls*
depression	*theory*
die	*engineering*
friends	*information*
suicide	*service*
depressed	*book*
suicidal	*legal*
feeling	*number*
mental	*center*
anxiety	*question*
hate	*phone*
pain	*engineer*
shit	*account*

We also analyze behavioral elements that explore posting frequencies in different time frames, and relational attributes. This work is based on Dataset 2 – suicidal ideation, which addresses Twitter data with 3 groups defined: suicidal ideation, focused control and random control.

4.3.2 Features explored

The features explored for this case are extracted from the Short Profile Version (SPV) of each user. To recall, the short profile version corresponds to a subset of relevant tweets related to suicidal ideation. It is given by the Short profile version classifier (SPVC) which detects individual tweets related to suicidal ideation. The tweets for which the SPVC provides a score over a given threshold (0.5) are retained as part of the SPV.

The features extracted are the following:

a) Generic Text-Based Features

These features address open vocabulary models such as bag of words models and word embeddings to represent text.

- Bag of Words and N-Grams

These are features that have been used to assess similar tasks, such as depression detection and eating disorders screening [91]. In our case, each user was represented by a

document consisting of the concatenation of the text of all their tweets. Afterward, we used the Scikit-learn [115] Python library: TfIdfVectorizer to generate a TF-IDF representation of {1-5}-grams at the word/term level. A set of Spanish stop words were considered to build this representation [141]. These features are referred to as BoW features in further sections.

We also used ekphrasis [18] as a text preprocessing tool to replace generic tag elements such as money, phone numbers, digits, hashtags, and emoticons. We also removed the n-grams that appeared in less than 5% of the documents to reduce the feature space.

- Word Embeddings

They are representations of textual terms as vectors of real numbers. Words that are semantically related have a similar representation over the vector space. The sequences of these representations are fed as inputs to train predictive models. These types of representations have been recently used in state-of-the-art approaches to address suicide risk assessment [39,146].

We made use of word embeddings previously learned over a dataset with 2 million Spanish tweets [43]. In this chapter we do not perform an analysis of these representations, but we do use them later to create the predictive models described in Section 5.3.

b) Behavioral and Psychological Features

These consist of a group of features based on generic lexicons [154], statistics measured from the users' writings, information of interest for clinicians regarding the behavior of users in time, the users' social network (relational features) [34], and lexicons, which include terms (n-grams) referring to suicidal ideation or suicide risk factors (we referred to these features as suicide-related lexicon features). They are also referred to as the social networks and psychological (SNPSY) features. Each of these types of features is described in the next items.

- Posting frequency

These features are based on the information extracted from the metadata of tweets. Here, we measured the behavior of users based on their activity within certain periods, which are defined at different granularity levels.

These features are detailed in Table 4.7. Some of these features take into account all the tweets of the user (full profile), while others are only extracted from the SPV.

Table 4.7. Description of posting frequency based features.

Feature	Description	Source
Working week tweets count ratio	Total number of tweets on weekdays (Monday to Friday) normalized by the total amount of tweets	SPV tweets
Weekend tweets count ratio	Total number of tweets on weekend days (Saturday and Sunday) normalized by the total amount of tweets	SPV tweets
Median time between tweets	Median of the time (in seconds) that passes between the publication of each tweet	SPV tweets
Sleep time tweets ratio	Ratio of tweets posted during the inferred sleep period of the user	Full profile tweets
Normalized tweet count per quarter (4 features)	Number of tweets posted by the user within each quarter of the year, normalized by the total amount of tweets generated by the user during the year	SPV tweets

SPV: short profile version.

The intention of the sleep time tweets ratio (STTR) is to identify the differences between control users and users at risk regarding the periods of the day on which they post. Considering that our data collection is delimited by language but not by location, that the posting time provided for a tweet is in coordinated universal time and not the time of the user location, and that not enough information from our data was found to automatically identify the location of all the users, we defined an approach to address this issue.

As explained in Equation 4.1, a day was divided into 8 fixed time slots of 3 hours each. Afterward, we assumed that an average user had at least around 6 hours of sleep time, and within this 6-hour period, a smaller number of tweets would be created in comparison to the rest of the day, so we counted the number of tweets (t) created within each 3-hour time slot for all the tweets of the full profile of a user. Next, for each user, we calculated the sum of the number of tweets within each pair of continuous time slots and selected the minimum score obtained by all the pairs. We also assumed that the first and last slots can be continuous. Finally, this value was normalized according to the total number of tweets of the full profile of the user (T). This feature was named as STTR:

$$STTR = \frac{min_{i=0\ldots7}\left\{t_i + t_{(i+1)mod\ 8}\right\}}{T} \quad (4.1)$$

It is important to recall that for the measurements that refer to a bigger granularity such as weekdays, weekends, and months, the impact of time difference is not as big as for features based on day periods.

- Tweets' Statistics

This group refers to 5 types of features that correspond to statistical measures calculated from the tweets of users. We considered elements such as the number of tweets created and their length and the number of tweets that were retained for each user at the SPV in relation to the total number of tweets posted. These features are described in Table 4.8.

Table 4.8. Description of tweets' statistics features.

Feature	Description	Source
Suicide-related tweets ratio	Ratio of tweets retained by the SPVC over all the tweets of the full profile	SPV and full profile tweets
Median SPVC score	Median of the scores obtained by the tweets that are part of the SPV after applying the SPVC	SPV tweets
Median tweet length	Median length of all the user tweets (word level)	SPV tweets
Number of SPV tweets	Number of tweets	SPV tweets
Number of user tweets	Number of tweets posted by the user since the creation of the account	Tweet metadata

SPVC: short profile version classifier.
SPV: short profile version.

- Relational Features

These are informative features regarding the relationships and interactions between users. Elements such as the count of retweets and favorites received and given by the users can provide insight on the social support they have, along with information regarding the number of followers and followees, as previously considered for depression screening [34]. Table 4.9 describes the relational features extracted for our evaluation.

Table 4.9. Description of relational features.

Feature	Description	Source
Followers number	Number of followers	Tweet metadata
Friends number	Number of accounts followed by the user	Tweet metadata
Favorites given	Total number of favorites given by the user	Tweet metadata
Median favorites count	Median of the favorites received by the user	SPV tweets
Median retweets count	Median of the retweets received by the user	SPV tweets

SPV: short profile version.

- Lexicons and Suicide Risk Factors Vocabulary

The use of lexicons has been proven to be successful for tasks dedicated to screen mental disorders [63]. For our approach, we counted the frequency of words belonging to all the categories of the Linguistic Inquiry and Word Count (LIWC) 2007 Spanish dictionary [130,154] normalized by the size (in number of terms) of the concatenated writings of the users.

To the dictionary, a group of other categories was added containing vocabulary and up to 3-gram phrases that could be mapped to suicide-related terms and risk factors such as suicide methods; terms referring to self-injuries; explicit suicidal ideation references; self-loathing terms; words that might imply disdain, insomnia, and fear; and possible references to previous suicide attempts, suffering from racial or sexual discrimination, eating disorders, substance abuse, bullying, lack of social support, and family and money issues, along with vocabulary that might imply that some sort of discrimination or abuse has been suffered, that someone close has died from suicide, and even vocabulary regarding the lack of spiritual beliefs, as religion is considered to be a protective factor for screening tasks [57].

The terms and phrases selected for these categories were based on manually mapping common terms and phrases seen in a sample of tweets labeled as suicide related during the dataset's first labeling process with the assessment of a clinician. These features were calculated using the SPV.

- Sentiment Analysis

We obtained a score for each tweet in terms of its polarity. For this purpose, we used senti-py [73], trained on Spanish texts from different sources, including Twitter. It is based on a BoW model with an intermediate feature selection process. To obtain a score per user, we calculated the median of the scores of all the tweets from the SPV.

c) Image-Based Feature

We used the output of a pre-trained classification model dedicated to the detection of images related to anorexia. The model was applied to each of the images extracted from the users' tweets of our dataset. To obtain a single score per user (images user score), the average of the individual scores of the images of each user was considered as the user's aggregated score.

4.3.3 Comparative analysis of groups

We performed an analysis of the features extracted to identify significant differences between the samples of users at risk and the control groups. For each feature extracted, we conducted an independent 2-sample Mann-Whitney U test among the suicidal ideation group of users and the different control groups. We also conducted this test to compare both of the control groups (focused and generic control groups). We performed a nonparametric test considering that our features do not follow a normal distribution and that there was no homogeneity of variance for most of them.

When comparing the suicidal ideation and focused control groups at the SNPSY features, we found significant differences with $p<.001$ among the following features: suicide related tweets' ratio, median time between tweets, verbs, verbs conjugated in singular of the first person ("I"+verb), cognitive mechanisms, anxiety-related terms, usage of personal pronouns, usage of the pronoun "I," negations, terms to express feelings, and cursing terms. Regarding suicide-related lexicons, the usage of suicide explicit terms, depression-related terms, self-loathing, substance abuse, self-injuries, and terms expressing lack of social support also presented an important significance ($p<.001$). Regarding the features from the BoW model, after conducting the same test, we found

significant differences with p<.001 for n-grams such as *I feel, sad, kill myself, cry/crying, depression, die (morirme), horrible, anxiety, die, pills*, among others.

Considering all the features used (24,758), a total of 522 features were significant for distinguishing the groups according to these tests with p<.001. Table 4.10 shows the medians and the distributions overlapping index [114] for both groups on a sample of relevant features.

Table 4.10. Medians and Distribution Overlapping Index for some of the attributes with the most significant differences between the Suicidal ideation and Focused control groups.

Attribute	Suicidal ideation median	Focused control median	Overlapping index
Anxiety	10.94	0	0.25
Coursing terms	21.52	7.68	0.43
Die (self-reference "morirme")	5.45	0	0.25
I feel	46.25	6.71	0.32
Self-loathing	0.03	0	0.35
Verb I (verbs conjugated in first person - singular)	22.66	12.11	0.41

When repeating the independent two-sample Mann-Whitney U test to compare the suicidal ideation risk group with the generic control set of users regarding the SNPSY features, among the ones with p<.001, we found that the median SPVC score, the number of tweets generated, and the median time between tweets were different among both groups (suicidal ideation risk vs generic control).

We identified differences in discussion topics such as money and work, about which the generic control users seem to discuss more, whereas the members of the suicidal ideation risk group use terms more related to health and biological aspects.

As in the previous case, the use of self-references was higher in the suicidal ideation risk group. Within the significant n-grams from the BoW model, we found terms such as *feel, to die, songs, someone, cry/crying, anxiety, life, breath, bad, and fear*.

Table 4.11 displays the median value and overlapping index of the distributions of the groups in terms of some of the attributes mentioned. Again, taking into account all the features used (24,449), 3,250 were significant for distinguishing between the suicidal ideation risk and the generic control group in terms of this test with p<.001.

Table 4.11. Medians and Overlapping Index for some of the attributes with the most significant differences between the Suicidal ideation and Generic control groups.

Attribute	Suicidal ideation median	Focused control median	Overlapping index
Median classifier score	0.72	0.65	0.46
To die (“morir”)	19.5	0	0.25
Number of tweets	2,076.5	453	0.38
Health terms	17.19	8.18	0.44
I (singular first person personal pronoun)	35.46	49.59	0.23
Work	41.32	9.60	0.44

Regarding other features explored, considering a 95% CI, for the suicidal ideation risk vs focused control groups, the number of friends (p=.04) and median tweet length (p=.04) were significantly different. For these cases, the median number of friends for a focused control user (578.5) was higher than the median number of friends at risk (372.0). The same was true for the median tweet length, based on the SPV, which was higher for focused control users with 16 words against 13 of the suicidal ideation risk users.

In addition, there were significant differences in the STTR (p=.049) and weekday count ratio (p=.01). Under the same CI, for the suicidal ideation vs. focused control case, the weekday count ratio (p=.001), the STTR (p=.004), along with the number of followers (p=.05), and the total amount of favorites given (p=.006) showed significant differences. In this sense, generic control users appeared to tweet more on weekdays (Monday to Friday) as well as focused control users, whereas the opposite behavior was found for suicidal ideation risk users.

Regarding the median STTR, generic control users obtained an STTR value of 0.02, whereas users at risk

obtained an STTR value of 0.04, meaning that users at risk seemed to tweet more at night compared with the generic and focused control users as well.

The image scores were also significantly different according to the test with p=.002 for the comparison between the suicidal ideation risk and generic control groups, considering a 95% CI.

Curiously, for the comparison of the image scores between the suicidal ideation risk group and the focused control group, the test scores were different, with p=.05. This can be explained by the fact that users providing information or news about suicide make use of similar images, which characterize the condition, making it difficult to find a significant difference only judging by pictures.

As can be seen in Table 4.12, for both the control groups and the suicidal ideation risk group, the median image scores were slightly higher for the suicidal ideation risk group.

Table 4.12. Medians and Overlapping Index for the images score between the suicidal ideation, focused control and generic control classes.

Attribute	Group	Focused control median	Overlapping index
Images score	Suicidal ideation	0.24	0.64
	Focused control	0.23	
	Suicidal ideation	0.24	0.52
	Generic control	0.23	

Finally, to compare our control groups (focused and generic control groups), we performed the same test (Mann-Whitney U test) and found significant differences between some of these groups' features (n=181) with p<.001. Among these features, we found mainly suicide-related lexicons, such as suicide methods, suicide explicit terms, bullying, discrimination, and substance abuse–related terms.

We also found differences (p<.001) in other textual, relational, and behavioral attributes, such as the number of tweets, number of friends, number of followers, median favorites and retweet counts, suicide related tweets' ratio, polarity score, median time between tweets, and STTR, among others.

4.4 Characterization of anorexia nervosa

4.4.1 Introduction

Here we address our work dedicated to the characterization of anorexia nervosa on social platforms. The work described in this section corresponds to our findings reported in [125] and [124]. As for the characterization of suicidal ideation, we explore several multimodal and behavioral features. We also introduce a deep analysis of the stages towards recovery focusing on the shift of interest across these stages.

This work is based on Dataset 3 – anorexia nervosa, which addresses Twitter data of users at the early stages of anorexia, users in treatment, recovered users, focused control, and random (generic) control users. We also take into account the participation of volunteers at the last stages of treatment from anorexia nervosa.

Through our work we: 1) extract and infer several features that consider multiple elements: images, texts, relations among users, posting patterns, and demographic information. These features are generated to identify elements that characterize users with AN at different stages of the illness and recovery. We also determine the elements that distinguish these users from two types of control cases. 2) We perform a deep analysis of the images of users with AN and control users to detect whether differences between these groups can be identified on the basis of visual properties. 3) We further explore the social network of users with AN through the detection of communities and the analysis of topics of interest of the different types of users, along with those of their followees. 4) We also perform a deeper analysis of users at the contemplation stage as these users are relevant for the further development of social recommendation methods. 5) Finally, we analyze the types of users followed by people with anorexia nervosa to get an insight of the types of suggestions provided by Twitter's contact recommender system (who to follow).

4.4.2 Analysis of features

As part of our data collection process, we extracted, calculated, and inferred some features for performing the analyses required. For this purpose, we considered network clustering

and visualization algorithms; prebuilt machine learning models for sentiment analysis; and age range and gender detection tools, including models for the detection of objects in images. We also used external sources with lexicons to detect emotions, topics of interest, risk factors, and affective processes.

After verifying that our numerical features did not follow a normal distribution and that there was no homogeneity of variance for most of them we used nonparametric tests. We used Mann-Whitney U's test to check for differences between pairs of groups of interest. As we considered some categorical elements as well, such as age groups, we transformed them into Boolean representations to perform a two-sided proportion z test among the groups with these feature types, which is a test equivalent to the proportions chi-square (X^2) test [151].

We analyzed different perspectives and several features within them:

a) Content and shared interests' perspective:

Through this perspective we analyze the textual content shared by users in their tweets. We consider linguistic and psychological aspects through six categories. Some of these categories were based on a classification given by the LIWC Spanish dictionary. The remaining categories were defined by considering psychological aspects related to EDs, which were defined under the supervision of clinicians.

The categories analyzed were mainly the same described for our initial analysis of multiple mental disorders described in Section 4.2.2. These features categories are: linguistic elements (24 features); affective processes and emotions, including polarity measures (29 features); personal concerns and biological processes (12 features); topics of interest to the users (200 topics); vocabulary related to suicide risk factors; and anorexia-related vocabulary (9 features). For this last feature type, on the basis of the work of Arseniev *et al.* [11] we used the translated categories of terms related to AN and its symptoms. We also kept some of the terms in English, as they are also used by Spanish-speaking users. In addition to these categories, we added names of known laxatives in Spanish [45].

The topics of interest of a user are analyzed as we would like to know if there is a shift in the main interests of users through the recovery path. Given the Twitter context, we

perform a different extraction approach from the one described in section 4.2.2, where topics were only extracted at an individual post level.

To address user level data we take into account that the topics of interest of a user are given by the interests of their followees, the content they like (given by the tweets made by others and marked as favorites), and by the content posted by themselves.

For each user, we collected 1) a random sample of their own tweets (up to 500 texts), 2) a random sample of 200 tweets that they had liked during the same period, and 3) the profile descriptions (biographies) of up to 200 random followees of the user. These tweets and descriptions are relevant enough samples of texts that characterized the interests of a user.

An individual score with Empath was obtained for each text (tweet or description). Later, the final score for a topic for a given user was calculated by averaging the scores obtained by the topic on all the tweets considered.

It is important to mention that as Empath's categories are in English, we add a translation step before the Empath scores' calculation, using the Googletrans Python API [67] for this purpose. The amounts of tweets and descriptions defined for this approach are also based on the request limitations of the API of Twitter and the Googletrans API.

We also analyze whether the proportion of tweets related to AN changes significantly according to the recovery stage, as it is expected that users at the initial stages produce more tweets related to their condition. For this purpose, we first built and compared two models to detect, for each user, if each of their tweets are related to AN. Second, we calculated the median score obtained by the classifier for all user tweets.

Finally, we compared the median values for all users belonging to a group to measure the presence of AN tweets in each group. It is expected that users with AN have a median value significantly higher than users in the treatment, recovered and control groups.

We trained two classifiers to distinguish tweets of two classes: 1) anorexia related and 2) control. The instances of the anorexia-related class corresponded to the individual tweets belonging to the users labeled as AN cases (1,766 tweets). Later, an equivalent number of tweets was selected to

represent the control class; these tweets were randomly extracted using Twitter's Sample Tweets API.

The first classifier was trained over a BoW model with {1-3}-grams at a term level. For this purpose, we generated a TF-IDF representation with {1-3}-grams. We considered Spanish stop words and used ekphrasis to replace terms referring to money, hashtags, and emoticons with generic tags. We used a LR method and 10-fold cross validation.

For the second classifier, a deep learning approach was applied. The model was defined through a CNN architecture that has been previously applied to text classification tasks [82], including a similar task for suicide risk assessment on social media [146].

The same preprocessing approach as that used for the prior model was applied. To train this model, tweets were represented as sequences of terms, and these terms were represented by pre-learned word embeddings that were trained over tweets in Spanish [43]. Each tweet was considered as an instance, and its label (anorexia related or control) corresponded to the class assigned to the tweet.

For the CNN, the embedding sequence instances were given as the model input, where a task-oriented fine-tuning was performed, and we applied a filter window ({2, 3, 5} terms). We applied max pooling and passed the output to a sigmoid layer to generate the final output.

Furthermore, 75% (2,649/3,532) of the instances were selected for training purposes and the remaining 25% (883/3,532) for testing. Among the training instances (tweets), 69.98% (1,854/2,649) were selected for training the model and 30.01% (795/2,649) were considered for validation.

The results found for this perspective were the most relevant for characterizing AN users. This perspective explored the textual elements from multiple points of view, including linguistic and psychological factors that were particularly useful in distinguishing AN users from control groups. These elements were also important for comparing our control groups, which were thought to exclusively differ from each other through the use of anorexia-related terms.

For the majority of the features analyzed for these perspectives, we calculated their median values for each group among all its users.

The p-values were also obtained to compare among the following pairs of groups: AN vs. treatment, AN vs. recovered, AN vs. random control, AN vs. focused control, and random vs. focused control.

- Linguistic dimension

Results for selected features of the 24 linguistic dimension features explored are listed in Table 4.13. We observed many linguistic features that could distinguish AN users from both control groups.

Notably, the use of first-person singular verbs, and consequently first-person singular pronouns, characterized the posts of AN users, along with a high use of negations and a reduced use of articles. In contrast with this, recovered and control cases make more use of first person plural pronouns ($p<.001$). Even users in treatment make more use of these pronouns compared to AN users ($p=.04$).

In addition, there were more features with highly significant differences between the AN group and the focused control group (22/24, 92% of features) than between the AN group and the random control group (15/24, 62% of features). This can be explained by the fact that, as shown in our further analysis, a high percentage of focused control users were organizations (*e.g.,* news sites, nutrition, and medical centers), and their linguistic features were quite distinguishable from those of users with personal accounts. This can also be noticed on the elements that distinguish between random and focused control users, as more personal accounts were part of the random control group.

Regarding the differences between the AN group and users in treatment, we observed significant differences in the use of second-person and first-person plural pronouns, which suggests that there might be a change in their attention focus and a higher level of interaction and inclusion with other people. This pattern was even more evident among recovered users.

Table 4.13. Comparative results between groups - Linguistic dimensions (***p<.001, **p<.01, *p<.05).

Feature	Median values					p-values and significance level (Mann-Whitney U)				
	AN	TREAT	RECOV	RAND CON	FOC CON	AN – TREAT	AN – RECOV	AN – RAND CON	AN – FOC CON	RAND – FOC CON
1st Person singular verbs	22.2E-03	20.5E-03	17.9E-03	8.14E-03	6.52E-03	.089	<.001 ***	<.001 ***	<.001 ***	.002 **
1st Person Singular Pronouns	41.9E-03	41.5E-03	3.0.5E-03	10.1E-03	5.62E-03	.299	<.001 ***	<.001 ***	<.001 ***	<.001 ***
1st Person Plural Pronouns	1.83E-03	2.26E-03	3.75E-03	3.72E-03	3.46E-03	.041 *	<.001 ***	<.001 ***	<.001 ***	.224
Second person plural pronouns	2.60E-03	3.38E-03	4.69E-03	2.69E-03	3.12E-03	.004 **	<.001 ***	.283	<.001 ***	.002 **
Third person plural pronouns	9.27E-03	9.51E-03	10.6E-03	11.8E-03	10.1E-03	.308	.047 *	<.001 ***	.059	<.001 ***
Negations	2.66E-02	2.44E-02	2.55E-02	1.98E-02	1.42E-02	.097	.283	<.001 ***	<.001 ***	<.001 ***
Adverbs	5.24E-02	4.79E-02	4.83E-02	3.81E-02	2.75E-02	.010 *	.003 **	<.001 ***	<.001 ***	<.001 ***
Articles	5.67E-02	6.15E-02	6.69E-02	7.06E-02	6.92E-02	.031 *	<.001 ***	<.001 ***	<.001 ***	.460 *
Verbs	1.72E-01	1.63E-01	1.71E-01	1.42E-01	1.37E-01	.072	.100	<.001 ***	<.001 ***	.001 **
Total pronouns	1.98E-01	1.94E-01	1.94E-01	1.49E-01	1.22E-01	.215	.182	<.001 ***	<.001 ***	<.001 ***
Past verb tense	1.84E-02	1.88E-02	1.69E-02	1.39E-02	1.18E-02	.431	.114	<.001 ***	<.001 ***	<.001 ***
Present verb tense	1.27E-01	1.22E-01	1.22E-01	1.05E-01	1.00E-01	.128	.171	<.001 ***	<.001 ***	.001 **
Median tweet length	14.00	14.00	13.50	12.50	19.00	.448	.432	.001 **	<.001 ***	<.001 ***

- Affective processes and emotions

The results of selected features are described in Table 4.14. As for the linguistic dimensions, there were significant differences between the values of users of the AN and focused control groups.

Negative emotions are found mainly for AN and treatment users; this can be observed also on the expression of emotions such as sadness, disgust, and anger, which are significantly higher for AN users than for control users. Within this same comparison, users with AN use more swearing terms and vocabulary that express anxiety and thoughts on their feelings and perceptions.

We observe that there are a few attributes with significant differences between AN and treatment users. For joy and positive emotions (LIWC), the scores were significantly higher for treatment users, which might reflect an improvement in the mood of people as they recover from AN.

Regarding recovered users, we also observed the existence of less negative emotions and more positive emotions than AN users. In fact, the expressions of anxiety of recovered users were significantly lower than those of AN users. In addition, their high score on social processes and the highly significant values in comparison with AN users suggest an openness to more interactions with other people.

Finally, the differences between random and focused control users are mainly observed through the use of swearing terms, the expression of positive emotions, cause and effects, insight, and discrepancies. In this sense, focused control users seem to be more formal and analytic toward things, which meets the characteristics of accounts that represent organizations.

For all the groups analyzed, we can observe in Figure 4.4 a radar chart that expresses the median values for the eight basic emotions defined by the wheel of emotions by Plutchik. We observed the predominance of sadness over all the other emotions in AN users.

- Personal concerns and biological processes

The results obtained for selected features are listed in Table 4.15. We observe that most of these features are relevant for distinguishing control users from AN users.

Table 4.14. Comparative analysis among groups based on effective processes and emotions.

Feature	Median values					p-values and significance level (Mann-Whitney U)				
	AN	TREAT	RECOV	RAND CON	FOC CON	AN - TREAT	AN - RECOV	AN – RAND CON	AN – FOC CON	RAND – FOC CON
Swearing	15.4E-03	14..6E-03	12.7E-03	8.06E-03	4.33E-03	.479	.188	<.001 ***	<.001 ***	<.001 ***
Positive emotions	5.93E-02	6.23E-02	6.50E-02	6.58E-02	5.79E-02	.039 *	.032 *	<.001 ***	.144	<.001 ***
Negative emotions	6.88E-02	6.50E-02	6.02E-02	4.97E-02	4.44E-02	.239	.010 *	<.001 ***	<.001 ***	.014 *
Anxiety	11.2E-03	12.2E-03	8.91E-03	6.61E-03	6.42E-03	.143	.004 **	<.001 ***	<.001 ***	.356
Cause and effect	1.79E-02	1.72E-02	1.44E-02	1.35E-02	1.62E-02	.131	.034 *	<.001 ***	.112	<.001 ***
Insight	3.89E-02	3.73E-02	3.72E-02	3.16E-02	3.60E-02	.299	.205	<.001 ***	.003 **	<.001 ***
Discrepancies	3.99E-02	3.72E-02	3.94E-02	3.03E-02	2.54E-02	.040 *	.177	<.001 ***	<.001 ***	<.001 ***
Senses and perceptions	5.03E-02	5.01E-02	4.89E-02	3.70E-02	3.80E-02	.479	.241	<.001 ***	<.001 ***	.200
Feel	16.0E-03	17.0E-03	13.7E-03	8.80E-03	9.30E-03	.422	.011 *	<.001 ***	<.001 ***	.284
Social processes	1.21E-01	1.21E-01	1.40E-01	1.23E-01	1.14E-01	.334	<.001 ***	.054	.177	.012 *
Joy	1.44E-02	1.55E-02	1.50E-02	1.35E-02	1.38E-02	.017 *	.293	.038 *	.020 *	.418
Sadness	2.43E-02	2.53E-02	2.24E-02	1.85E-02	1.78E-02	.299	.023 *	<.001 ***	<.001 ***	.236
Disgust	1.58E-02	1.65E-02	1.41E-02	1.27E-02	1.10E-02	.482	.077	<.001 ***	<.001 ***	.003 **
Anger	1.51E-02	1.64E-02	1.42E-02	1.39E-02	1.23E-02	.109	.154	.004 **	<.001 ***	.022 *

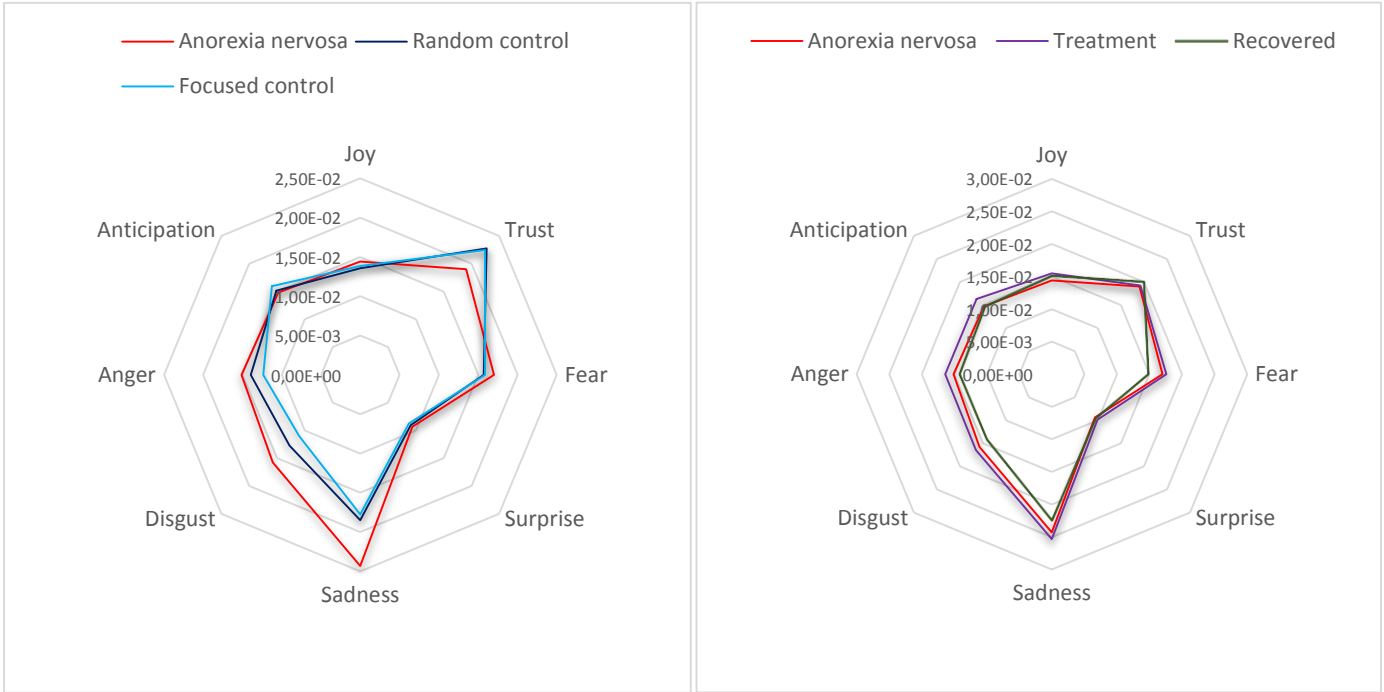

Figure 4.4. Comparative scores for emotions (left: AN and Control groups - right: AN, Treatment and Recovered groups) according to the basic emotions of Plutchik's wheel of emotions.

Table 4.15. Comparative analysis among groups based on personal concerns and biological processes.

Feature	Median values					p-values and significance level (Mann-Whitney U)				
	AN	TREAT	RECOV	RAND CON	FOC CON	AN – TREAT	AN - RECOV	AN – RAND CON	AN – FOC CON	RAND – FOC CON
Work	3.21E-02	3.50E-02	3.64E-02	4.95E-02	5.25E-02	.051	.013 *	<.001 ***	<.001 ***	.089
Achievement	3.87E-02	4.08E-02	3.92E-02	4.39E-02	4.30E-02	.069	.475	<.001 ***	<.001 ***	.181
Leisure	1.70E-02	1.91E-02	1.67E-02	1.99E-02	2.12E-02	.157	.399	<.001 ***	<.001 ***	.265
Money	8.22E-03	9.42E-03	11.7E-03	13.8E-03	12.2E-03	.033 *	.006 **	<.001 ***	<.001 ***	.003 **
Religion	2.16E-03	2.67E-03	3.42E-03	5.17E-03	3.38E-03	.102	<.001 ***	<.001 ***	<.001 ***	<.001 ***
Sexual	10.6E-03	11.6E-03	15.6E-03	7.98E-03	7.70E-03	.393	.019 *	.001 **	<.001 ***	.220
Death	10.5E-03	8.33E-03	6.50E-03	5.87E-03	6.24E-03	.047 *	<.001 ***	<.001 ***	<.001 ***	.245
Biological processes	9.02E-02	6.86E-02	6.41E-02	3.21E-02	5.01E-02	.035 *	.003 **	<.001 ***	<.001 ***	<.001 ***
Body	2.96E-02	2.45E-02	1.92E-02	1.21E-02	1.55E-02	.069	<.001 ***	<.001 ***	<.001 ***	<.001 ***
Ingest	33.8E-03	17.4E-03	15.3E-03	8.52E-03	11.2E-03	.011 *	.001 **	<.001 ***	<.001 ***	<.001 ***
Health	17.7E-03	17.5E-03	17.6E-03	6.83E-03	13.2E-03	.499	.467	<.001 ***	<.001 ***	<.001 ***

Control users discuss more about common concerns such as work, leisure, achievement, money, and religion, whereas AN users are more interested in aspects related to their image as seen for the categories body, ingest, health, and biological processes. There was also a significantly higher interest in death, compared with all the other stages.

For the treatment group, we observed significantly lower values for the ingest and biological process categories, which might be a sign of improvement in their condition compared with AN users. This is more evident in the comparison of AN and recovered users, where there are very significant differences among the same features. Note that the reference to religious aspects is lower for the AN, treatment, and recovered users in comparison with random control users.

Regarding random and focused control users, there are differences in the scores for the body, ingest, health, and biological process categories, as these are the ones that refer to signs of the illness.

Focused control users are characterized by their use of AN-related terms, and these findings suggest that among the focused control users, we can find people and organizations that often address the topic of AN. Among these, we can find foundations, medical centers, nutritionists, and psychologists. We later validated this assumption through a social network analysis.

- Risk factors' vocabulary

For the use of vocabulary related to risk factors, we noticed that a large number of features were highly significant for the comparison of the AN and control groups. In fact, all the risk factors considered were significant for distinguishing AN from random control users, as shown in Table 4.16. The use of suicide-related terms is higher for AN users than for all the other groups. Hate and self-loathing terms are found in a lower percentage for recovered users than for AN users and treatment users.

We observe that the use of terms related to bullying is higher for treatment and recovered users, which can be explained by the fact that while being on treatment and after recovery, patients are more likely to recognize the issues behind their ED.

Table 4.16. Comparative analysis among groups based on vocabulary related to risk factors.

Feature	Median values					p-values and significance level (Mann-Whitney U)				
	AN	TREAT	RECOV	RAND CON	FOC CON	AN – TREAT	AN – RECOV	AN – RAND CON	AN – FOC CON	RAND – FOC CON
Hate	98.4E-05	68.5E-05	33.1E-05	7.60E-05	2.90E-05	.146	<.001 ***	<.001 ***	<.001 ***	.028 *
Suicide related terms	4.20E-05	2.20E-05	0	0	0	.042 *	.002 **	<.001 ***	<.001 ***	<.001 ***
Self-harm	1.60E-05	4.60E-05	1.00E-05	0	0	.029 *	.209	<.001 ***	<.001 ***	.024 *
Work/school problems	8.80E-05	12.0E-05	6.80E-05	1.60E-05	3.10E-05	.412	.164	<.001 ***	<.001 ***	.116
Self-loathing	4.20E-05	1.90E-05	0	0	0	.247	.003 **	<.001 ***	<.001 ***	.007 **
Bullying	0	3.00E-06	11.0E-06	0	0	.059	.032 *	<.001 ***	.024 *	<.001 ***
Drugs or alcohol abuse	125E-06	143E-06	77.0E-06	6.00E-06	124E-06	.095	.299	<.001 ***	.401	<.001 ***
Lack of social support	0	2.00E-06	0	0	0	.252	.290	<.001 ***	<.001 ***	.010 *
Relationship issues	5.80E-05	7.00E-05	7.30E-05	0	1.50E-05	.138	.325	<.001 ***	<.001 ***	.001 **
Anti-depressants usage	0	0	0	0	0	.363	.343	<.001 ***	<.001 ***	<.001 ***

In general, the scores obtained by all the groups for these features are very low, as these are issues that do not seem to be openly addressed often.

- Anorexia-related vocabulary

These features address the use of vocabulary that describes certain signs and symptoms of AN. The results are presented in Table 4.17. All the features are highly significant for distinguishing AN users from control cases, and they are all highly significant for distinguishing random from focused control cases.

We observed that the scores obtained for the focused control cases were higher than the scores obtained for the random control users. This also happens for the case where recovered and AN users are compared; these users highly differ in the use of vocabulary dedicated to the promotion of AN and vocabulary that expresses concerns regarding body image, body weight, compensatory behavior, and laxatives references, along with caloric restrictions. AN users showed higher scores on these aspects.

We also observed that users in the treatment group had lower median values for almost all the features considered, with significant differences of up to four features in comparison with the AN group.

- Topics of interest

Here, we present the results for the exploration of the topics of interest of users that include anorexia-related terms in their texts (AN, treatment, recovered, and focused control users).

We assume that the interests of random control users are different and depend on the user. This is due to the fact that we do not consider common interest for these users during the data collection process.

Table 4.18 shows the top 20 topics of interest for the groups according to the Empath categories. We observe that, apart from the elements in common among groups, only users of the AN group refer to topics such as pain, eating, violence, and suffering.

Treatment users have many interests in common with AN users, but we can also observe other topics of interest such as reading, music, and sports.

Table 4.17 Comparative analysis among groups based on anorexia-related vocabulary.

Feature	Median values					p-values and significance level (Mann-Whitney U)				
	AN	TREAT	RECOV	RAND CON	FOC CON	AN – TREAT	AN – RECOV	AN – RAND CON	AN – FOC CON	RAND – FOC CON
Anorexia promotion	35.4E-04	23.5E-04	13.0E-04	4.99E-04	8.47E-04	.023 *	<.001 ***	<.001 ***	<.001 ***	<.001 ***
Body image	23.5E-04	7.01E-04	4.26E-04	0	1.60E-04	.010 *	<.001 ***	<.001 ***	<.001 ***	<.001 ***
Body weight	75.7E-05	32.9E-05	9.00E-05	0	8.90E-05	.106	<.001 ***	<.001 ***	<.001 ***	<.001 ***
Food and meals	29.5E-04	21.5E-04	16.8E-04	1.76E-04	5.20E-04	.159	.022 *	<.001 ***	<.001 ***	<.001 ***
Eat verb	222E-06	100E-06	91.0E-06	0	9.00E-06	.014 *	.001 **	<.001 ***	<.001 ***	<.001 ***
Caloric restriction	443E-06	34.0E-06	2.00E-06	0	0	.001 **	<.001 ***	<.001 ***	<.001 ***	<.001 ***
Binge eating	3.10E-05	3.40E-05	0	0	0	.400	.004 **	<.001 ***	<.001 ***	<.001 ***
Compensatory behavior and laxatives	9.00E-04	4.88E-04	2.71E-04	0	0	.090	<.001 ***	<.001 ***	<.001 ***	<.001 ***
Exercise	164E-05	91.8E-05	45.2E-05	7.30E-05	43.7E-05	.050	.001 **	<.001 ***	<.001 ***	<.001 ***

Table 4.18. Top 20 topics of interest (using Empath) among groups that use anorexia-related vocabulary and their median values.

AN		Treatment		Recovered		Focused control	
Topic	Median	Topic	Median	Topic	Median	Topic	Median
negative emotion	15.86	negative emotion	10.50	negative emotion	10.85	health	9.39
friends	8.24	friends	7.28	friends	9.78	communication	7.41
speaking	7.67	positive emotion	6.95	speaking	8.09	business	6.90
positive emotion	7.27	speaking	6.74	positive emotion	7.60	work	5.98
children	6.41	Social media	6.62	communication	7.01	positive emotion	5.88
pain	6.29	children	6.55	children	6.66	internet	5.83
eating	6.19	communication	6.06	family	6.46	negative emotion	5.82
communication	6.13	optimism	5.61	social media	5.62	social media	5.71
optimism	5.93	family	5.23	home	4.95	speaking	5.54
family	5.91	party	4.85	party	4.95	sports	5.44
love	5.60	love	4.80	optimism	4.91	messaging	5.03
shame	5.40	reading	4.51	love	4.91	college	5.01
violence	5.21	music	4.22	eating	4.31	eating	4.92
party	4.84	home	4.10	wedding	3.95	children	4.86
Social media	4.70	internet	4.09	sports	3.94	school	4.84
suffering	4.25	musical	3.99	giving	3.94	family	4.59
home	4.24	listen	3.94	violence	3.89	reading	4.41
hate	4.08	wedding	3.79	childish	3.75	party	4.30
childish	4.06	violence	3.63	pain	3.75	optimism	4.28
feminine	3.87	sports	3.62	affection	3.71	meeting	4.26

Similarly, recovered users rank topics such as sports and weddings in their list. Focused control users also express interest on different topics, with the highest scored topics being health, communication, business, work, internet, and sports, which matches with our prior assumptions regarding this group. Note that for visualization purposes, the actual median values were multiplied by 1000.

To explore the topics of interest in which the groups differed the most from the AN group, we performed the Mann-Whitney U test.

Figure 4.5 shows the top 20 topics with the most significantly different values ($p<.05$) between the AN group and the focused control, treatment, and recovered groups.

We observe that swearing terms, feminine terms, hate, pain, and appearance obtained high scores for AN users, whereas topics such as economics, college, photography, and work obtained high scores for focused control users.

We also observed a limited interest in topics such as music and art for AN users, in comparison with users in treatment, whereas these users (treatment) are less concerned about body and exercise in comparison with users from the AN group. Recovered users are also more concerned about general topics such as law, crime, and politics in comparison with AN users.

We report on the percentage of topics found with significant differences among the values for each group ($p<.05$): AN versus focused control, 61% (122/200); AN versus recovered, 40% (80/200); and AN versus treatment, 46.5% (93/200). We also calculated the values for Spearman rank correlation coefficient based on the median values obtained for each topic in each group.

The following pairs of groups were compared: AN versus recovered ($\rho=0.87$), AN versus treatment ($\rho=0.87$), AN versus focused control ($\rho=0.67$), treatment versus focused control ($\rho=0.87$), recovered versus focused control ($\rho=0.87$), and treatment versus recovered ($\rho=0.97$).

We observe that AN and focused control users are less interested in similar topics, whereas treatment and recovered users' interests are more correlated with those of the focused control group.

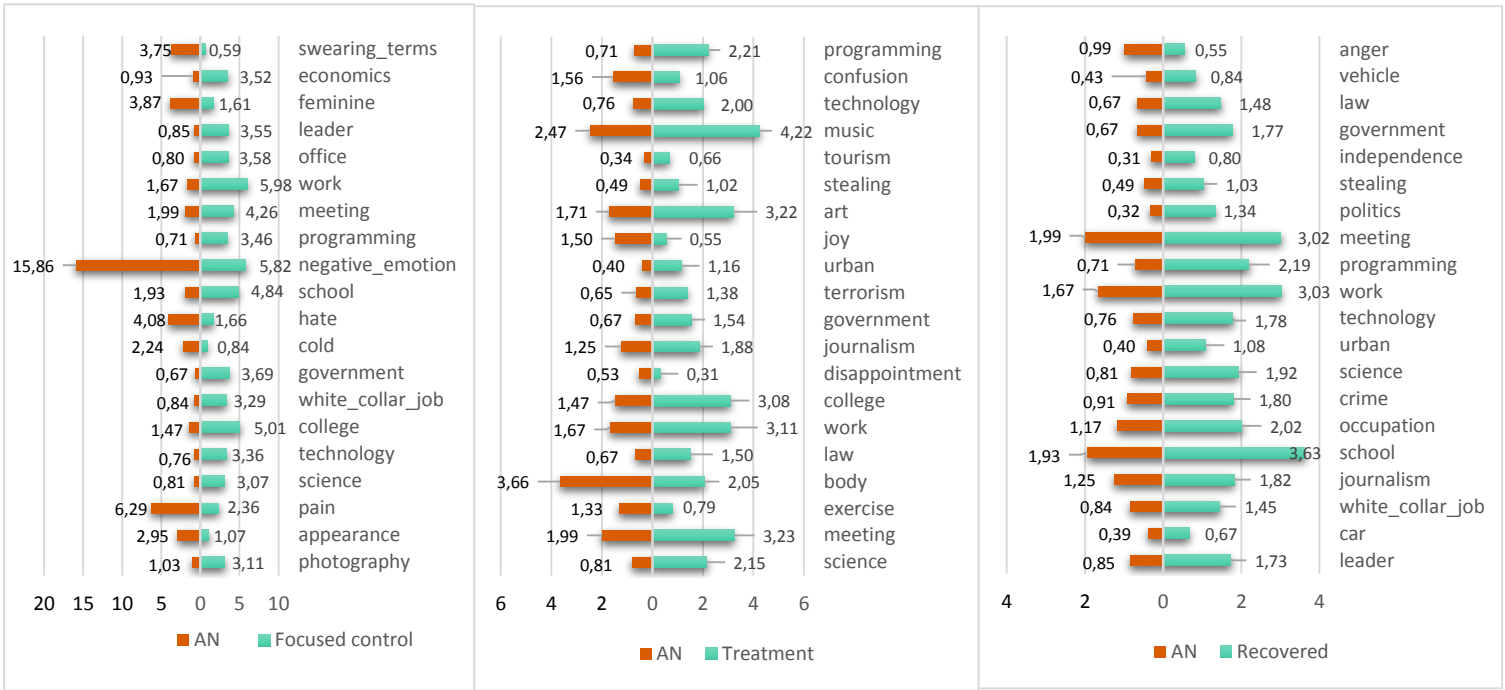

Figure 4.5. Top 20 topics with most significantly different values (p<.05) between the AN group and the focused control, treatment, and recovered groups respectively. The median values for each feature are shown.

- Proportion of AN-related tweets

The results obtained by the LR classifier at the training approach (cross validation) were as follows: F1 score=0.97, precision=0.98, and recall=0.97, whereas the results for the deep learning approach averaged after multiple runs over the test set were as follows: precision=0.98, recall=0.98, and F1 score=0.98.

As the second model obtained slightly better results, it was applied to all the tweets of all the users regardless of the group they belonged to. For each user, the value considered as a feature was the median score obtained by the classifier on all tweets. We then compared the median values of each group analyzed.

We used the Mann-Whitney U test to perform an analysis of the median score provided by the classifier to all the users' tweets. We applied the classifier to all groups of users. The median values for each group are the following: AN (0.23), treatment (0.13), recovered (0.08), random control (0.03), and focused control (0.05). The p-values for the group comparisons are the following: AN versus recovered ($p<.001$), AN versus treatment ($p=.004$), AN versus focused control ($p<.001$), treatment versus focused control ($p<.001$), and focused control versus random control ($p=.02$).

We noticed very significant differences between the AN group and all other groups considered. Notably, the median classifier score obtained by AN users was higher than that obtained by users from all other classes. Moreover, the median values for the groups decreased according to the recovery stage, meaning that the score was lower for recovered users than for treatment users. Note that focused control users obtain a higher score than random control users, as focused control users address AN-related topics.

b) Social network

Features are extracted taking into account the social network of the user. We analyze some features that characterize the user´s popularity and the support received by other users. These features correspond to the number of followers, favorites, and retweets of their posts. We focused on the social network (followees) of users that make use of anorexia-related terms, as our goal was to detect communities among these types of users. Furthermore, we explored the

likelihood of users with AN to follow users living with the disorder or anorexia promoters by analyzing the topics of interest of their followees. We also explored the differences in their interests and those of the followees of users in treatment and the followees of recovered users. The elements analyzed for this perspective are:

- Measures of interactions and engagement

These features are extracted from the metadata of the users' tweets. These features tell us about the relationships and interactions of AN users, which can differ from the interactions of control users. The features extracted and calculated for each user are as follows: number of followees, number of followers, total number of favorites given to the posts of other users, median number of favorites received by the user, and median number of retweets received by the user. These last two features were calculated by considering the user's full profile.

The results obtained for these features (Table 4.19) show that focused control users have a significantly higher median number of followers and followees than AN users. The median number of followers of these users (focused control) shows that these accounts have a higher number of followers than random control users, which might be an indicator of the popularity of these user types that are more likely to be organizations. We also observe that AN users have a reduced number of interactions with other users in comparison with treatment, recovered, and random control users (based on the favorites given). In general, we observe that a reduced number of tweets generated by all user groups are liked or retweeted by other users, probably because they consume this type of information in a discrete way or because they do not generate very popular content.

- Analysis of followees and communities' detection

We explored the network of users that made use of anorexia-related terms, corresponding to the AN, treatment, recovered, and focused control groups. This was done with the purpose of identifying characteristics of the network that were capable of distinguishing the groups defined, in particular the AN group and the focused control group, as users representing organizations that provide medical and psychological support

Table 4.19. Comparative analysis among groups based on interaction and engagement measures.

Feature	Median values					p-values and significance level (Mann-Whitney U)				
	AN	**TREAT**	**RECOV**	**RAND CON**	**FOC CON**	**AN - TREAT**	**AN - RECOV**	**AN – RAND CON**	**AN – FOC CON**	**RAND – FOC CON**
Number of followers	621.50	815.00	600.00	540.00	1,174.00	.017 *	.256	.256	<.001 ***	<.001 ***
Number of followees	286.50	483.50	289.50	492.00	509.00	.022 *	.177	<.001 ***	<.001 ***	.241
Given favorites	7,746.50	10,893.00	23,955.00	10,085.50	4,917.00	.035 *	.004 **	.019 *	.018 *	<.001 ***
Received favorites	0.00	1.00	0.50	0.00	1.00	.110	.219	.001 **	.001 **	<.001 ***
Received retweets	0.00	0.00	0.00	0.00	0.00	.043 *	.286	.067	<.001 ***	<.001 ***

could be part of it, and it would be relevant to get an insight into the relationships between both groups. For this purpose, we extracted a sample of up to 100 followees of each user from each of these groups (considering Twitter's API request limitations).

We built a directed graph where a link between two nodes was given by a follows relationship, meaning that users, represented by nodes, are linked to other nodes through directed edges where the arrowheads point to the users they follow.

Later, a clustering algorithm was applied to detect communities among these users. We then performed a comparison between the communities automatically detected and what we defined as validation groups, which were created considering the followees of the AN, treatment, recovered, and focused control groups. These validation groups were defined in such a way that a user was assigned to a validation group (AN, treatment, recovery, or control) if it was mostly followed by users belonging to the originally labeled groups. This was done taking into account up to two followees' levels, denoted as validation subgroups, as explained in Table 4.20, where we describe the general organization of a group.

Table 4.20. Groups for social network analysis based on users' labels.

Group	Subgroup	Nodes - user type
Group X	G_i	Users manually labeled as part of the X group
	G_{i+1}	Users mostly followed by G_i
	G_{i+2}	Users mostly followed by G_{i+1}

We considered four main groups and three subgroups per group, where the first subgroup always corresponded to the original users labeled. An instance of a validation group would be Group AN, which is composed of three subgroups: G1 composed of the originally labeled AN's users, G2 composed of the users mostly followed by G1 users, and G3 composed mostly of users followed by G2.

On the basis of a manual revision of a sample (translated to English) of the profile descriptions of users belonging to the communities detected with most nodes, we performed a further

analysis of the types of users that were identified as part of each community, and we mapped these communities to our predefined groups so that we could identify which type of users from our groups of interest were part of the communities detected. For the visualization of the social network, we considered the Force Atlas 2 [77] algorithm, and for the detection of communities, we used Louvain's method [20]. Both of these methods are implemented on Gephi [17].

In Table 4.21, we report on the percentages of nodes belonging to each group defined through the approach previously explained in Table 4.20. Most of the users considered were part of the focused control group, followed by AN, recovered, and treatment users.

For visualization of these groups, we used Gephi, as shown in Figure 4.6. A total of 99,283 nodes were considered, with each node representing a user. The average number of edges per node (average degree of the graph) was 2.57, the shortest distance between the two most distant nodes in the network (full network diameter) was 15, and the average path length was 4.72, which represents the average number of steps it takes to get from one member of the network to another. The average clustering coefficient was 0.017, which implies that most of the nodes were not related.

To ease the visualization and interpretation of the results, we applied a k-core filter with k=2 to see the maximal subgraph with a minimum degree equivalent to k. The number of nodes displayed in Figure 4.6 is 12,680, and the size of the nodes is given by the page rank score obtained by each node. The graph clearly shows the polarization between the AN and focused control groups, with the few treatment and recovery cases displayed in between and closer to the focused control cases.

For further analysis of the full network, we applied a clustering algorithm to detect the communities within it. We found 80 communities and obtained a modularity value of 0.86. As shown in Table 4.22, we analyzed the descriptions (biographies) of the users of the 10 communities with the highest node percentages.

We describe the types of users found in each community and identify the types of users from our annotated groups that are part of each community.

Table 4.21. Graph information of sub-groups defined based on their followers' type.

Validation Group	Sub-Group	Nodes - user type	Nodes percentage
AN	G_1	AN	0.05%
	G_2	Mostly followed by G_1	10.55%
	G_3	Mostly followed by G_2	12.46%
Focused control	G_4	Focused Control	0.16%
	G_5	Mostly followed by G_4	46.56%
	G_6	Mostly followed G_5	20.57%
Treatment	G_7	Treatment	0.01%
	G_8	Mostly followed by G_7	3.95%
	G_9	Mostly followed G_8	0.74%
Recovered	G_{10}	Recovered	0.02%
	G_{11}	Mostly followed by G_{10}	4.63%
	G_{12}	Mostly followed G_{11}	0.30%

Figure 4.7 shows the network, with the automatically identified communities highlighted. For comparison, we used the same structure displayed in Figure 4.6.

It can be seen that the community with the highest number of nodes is GC1, which corresponds to the community of users that are likely to have an ED and users that might be anorexia and bulimia promoters (they correspond to big nodes in the graph, i.e., higher page rank, meaning more popular nodes).

We also observe two other relevant communities that mainly correspond to focused control cases (GC2 and GC3) and are characterized by having users that represent organizations and specialists on mental health issues and nutrition centers.

We also observe a community that corresponds to news and TV accounts (GC5), which also characterizes focused control users.

We see that the small number of treatment and recovered users are part of different communities that address multiple topics and that the communities that gather users from different groups are those that involve singers, artists, influencers, and leisure-related topics.

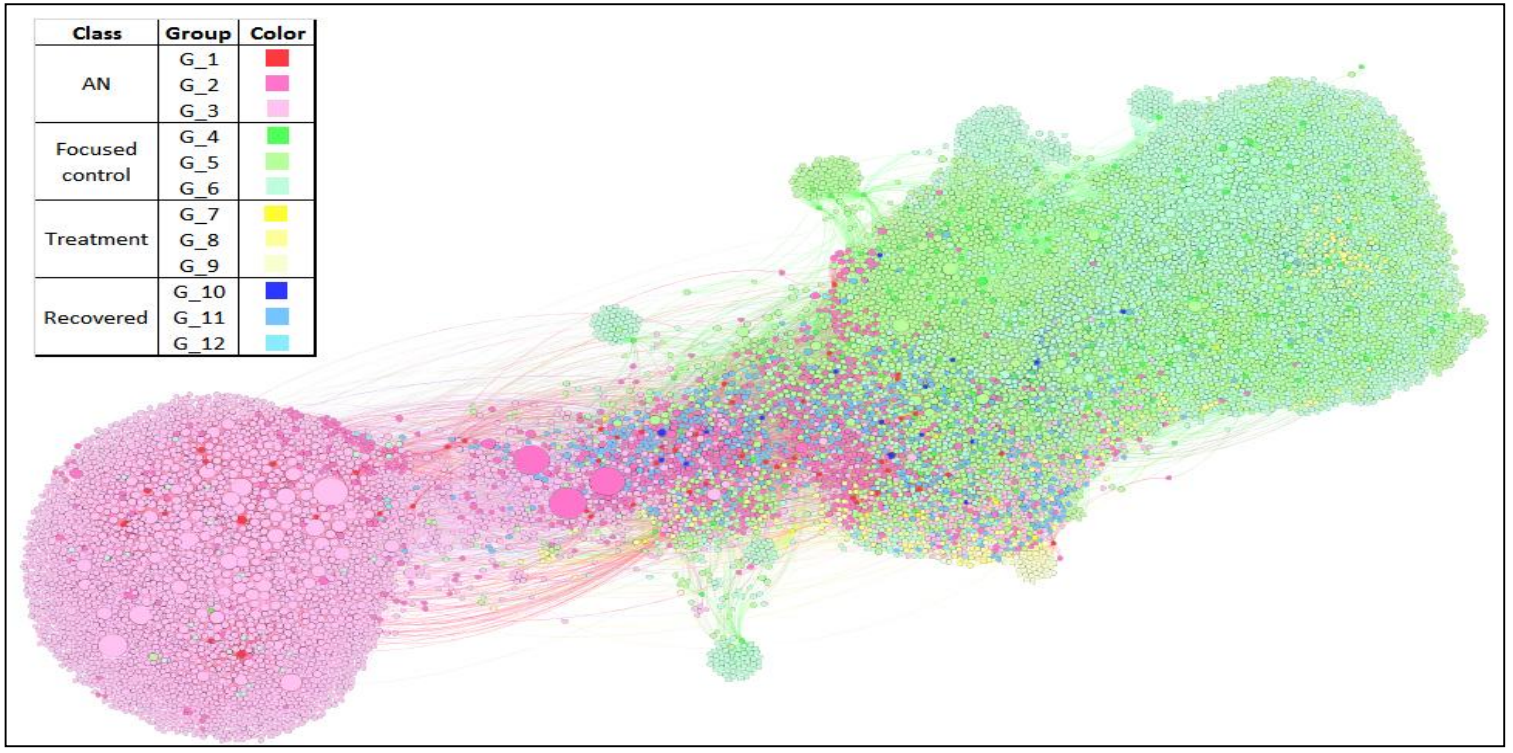

Figure 4.6. Visualization of the social network of the AN, focused control, treatment, and recovered groups according to the types of users they are mostly followed by. Each group is represented by a different color. Groups associated with the same class have similar colors. AN: anorexia nervosa; G: group ID.

Table 4.22. Description of the types of users identified in each community with the highest node percentages.

Community	Community description	Group of users identified	Nodes percentage
GC_1	Users with Eating disorders, and anorexia and bulimia promoters	AN	9.65%
GC_2	Organizations, medical centers and psychologists	Focused Control	8.04%
GC_3	Nutritionists, nutrition centers	Focused Control	7.37%
GC_4	Varieties - influencers	Focused Control	3.88%
GC_5	News and TV	Focused Control	3.72%
GC_6	Pop singers' fans	AN and Recovered	2.69%
GC_7	Undefined varieties	Treatment and Focused Control	2.54%
GC_8	Undefined varieties	Recovered and AN	2.53%
GC_9	Comics, anime, drawing	Treatment and Focused Control	2.31%
GC_{10}	Uruguay community	Focused Control	2.29%

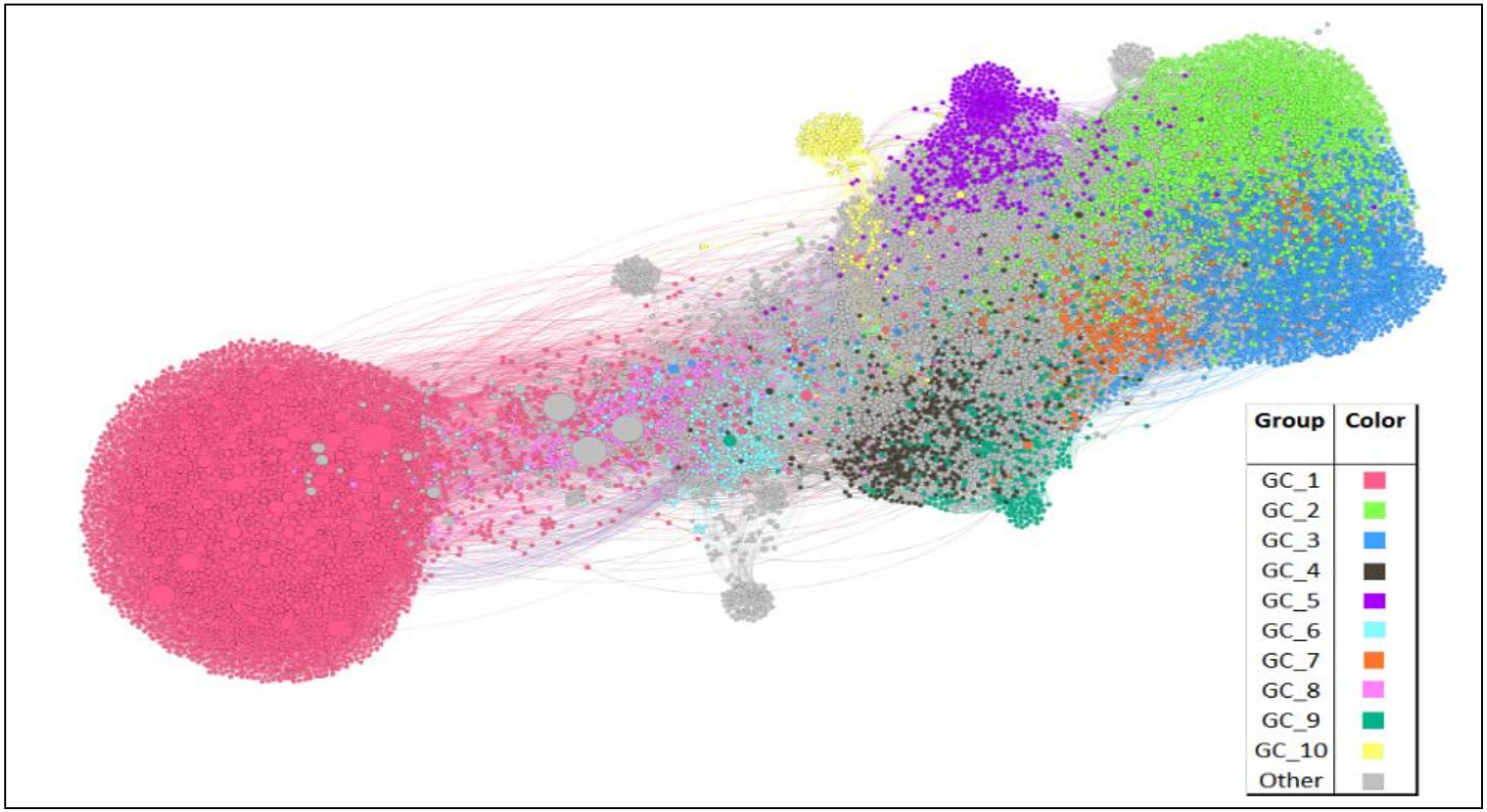

Figure 4.7. Graph visualization of the 10 communities detected with the highest node's percentages.

These results show that users at the precontemplation and contemplation stages are isolated from accounts that offer assistance to overcome the illness. In this sense, recommender systems might enforce this behavior of the network because they tend to recommend a user to follow similar accounts.

- Analysis of interests between users and their followees

As it is our purpose to identify the topics of interest of AN users' followees, we follow the process applied for the analysis of the topics of interest of users of each group, but in this case, we address the followees of each user type. As shown in Figure 4.8, for this case, we considered up to 25 followees from a sample of up to 25 users per group analyzed. Then, for each of these followees, we calculated scores for the Empath topics by considering the descriptions of 25 random followees, a random sample of their own tweets (up to 200 texts), and a random sample of 200 tweets that they had liked during the same period.

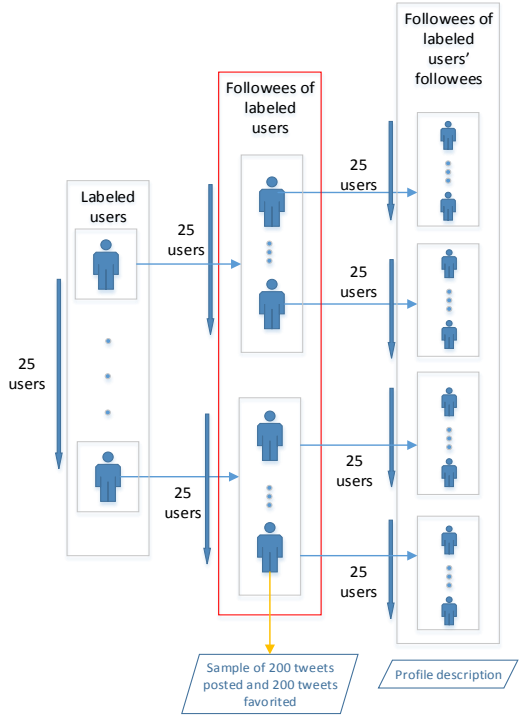

Figure 4.8. Structure defined for the extraction of the interests of the followees of a given user group. For each labeled user of a group, we analyzed the tweets posted and liked by their followees and the profile description of the followees of the labeled users'.

The score of a topic for each followee of a user is given again by the average score obtained from all the texts considered. The score for a topic of a user is given by the median of the scores of their followees. Once the scores for the samples of users representing each group were obtained, we calculated the median value of each topic using the scores of each user belonging to the group. Later, we performed a comparison between the interests of users of each group (calculated before) and those of their followees.

Regarding the results for the topics of interest of the users who make use of anorexia-related terms, Table 4.23 shows the top 20 topics of interest for our groups' followees according to the Empath categories. We observe that negative emotions, eating, pain, death, and violence are among the topics most relevant to AN followees.

Regarding the other groups, we cannot observe a pattern that would normally characterize each user type; instead, we observe interest in all types of topics, which is more evident in focused control users. We can also observe that topics such as friends, family, children, and parties are relevant for most of the groups. For a better comprehension of the results on this topic analysis task, we explored the topics in which certain followee groups differ the most. We used the Mann-Whitney U test for this purpose.

Figure 4.9 shows the top 20 topics, with the most significantly different values ($p<.05$) between the AN followees group and the focused control, treatment, and recovered followee groups.

There is a very high value for negative emotion on AN followees in comparison with focused control followees. Appearance is also a topic in which AN followees differ from focused control followees and recovered followees.

We also report on the percentage of topics found with significant differences among the median values for the following pairs of groups ($p<.05$): AN followees and recovered followees, 45% (90/200); AN followees and focused control followees, 75% (150/200); AN followees and treatment followees, 48% (96/200); AN and AN followees, 22% (44/200); and recovered and recovered followees, 21% (42/200).

Table 4.23. Top 20 topics of interest and their Empath median values for the groups´ that make use of anorexia related vocabulary followees.

An followees		Treatment followees		Recovered followees		Focused control followees	
Topic	**Median value**	**Topic**	**Median value**	**Topic**	**Median value**	**Topic**	**Median value**
negative emotion	10.86	negative emotion	7.92	negative emotion	9.73	business	7.29
friends	7.58	friends	7.76	friends	7.71	communication	7.04
Positive emotion	7.30	Positive emotion	7.02	Positive emotion	7.23	work	6.80
speaking	6.50	communication	6.96	communication	7.15	Positive emotion	6.41
communication	5.93	Social media	6.83	speaking	6.87	internet	5.72
optimism	5.85	speaking	6.42	Social media	5.69	Social media	5.36
children	5.80	children	5.48	optimism	5.55	party	4.86
Social media	5.71	party	5.43	party	5.22	meeting	4.85
party	5.62	optimism	5.22	children	5.20	speaking	4.79
love	5.09	family	4.82	family	4.98	negative emotion	4.66
family	4.69	love	4.28	internet	4.33	leader	4.65
childish	4.03	internet	4.15	giving	4.00	reading	4.60
giving	4.00	music	4.09	love	3.96	school	4.58
eating	3.93	messaging	4.06	messaging	3.83	messaging	4.57
home	3.83	listen	4.02	reading	3.73	children	4.56
pain	3.72	musical	3.92	wedding	3.70	occupation	4.50
death	3.72	reading	3.91	celebration	3.67	family	4.49
wedding	3.70	wedding	3.76	listen	3.57	optimism	4.35
violence	3.62	celebration	3.74	home	3.52	government	4.31
celebration	3.42	childish	3.72	childish	3.47	celebration	4.12

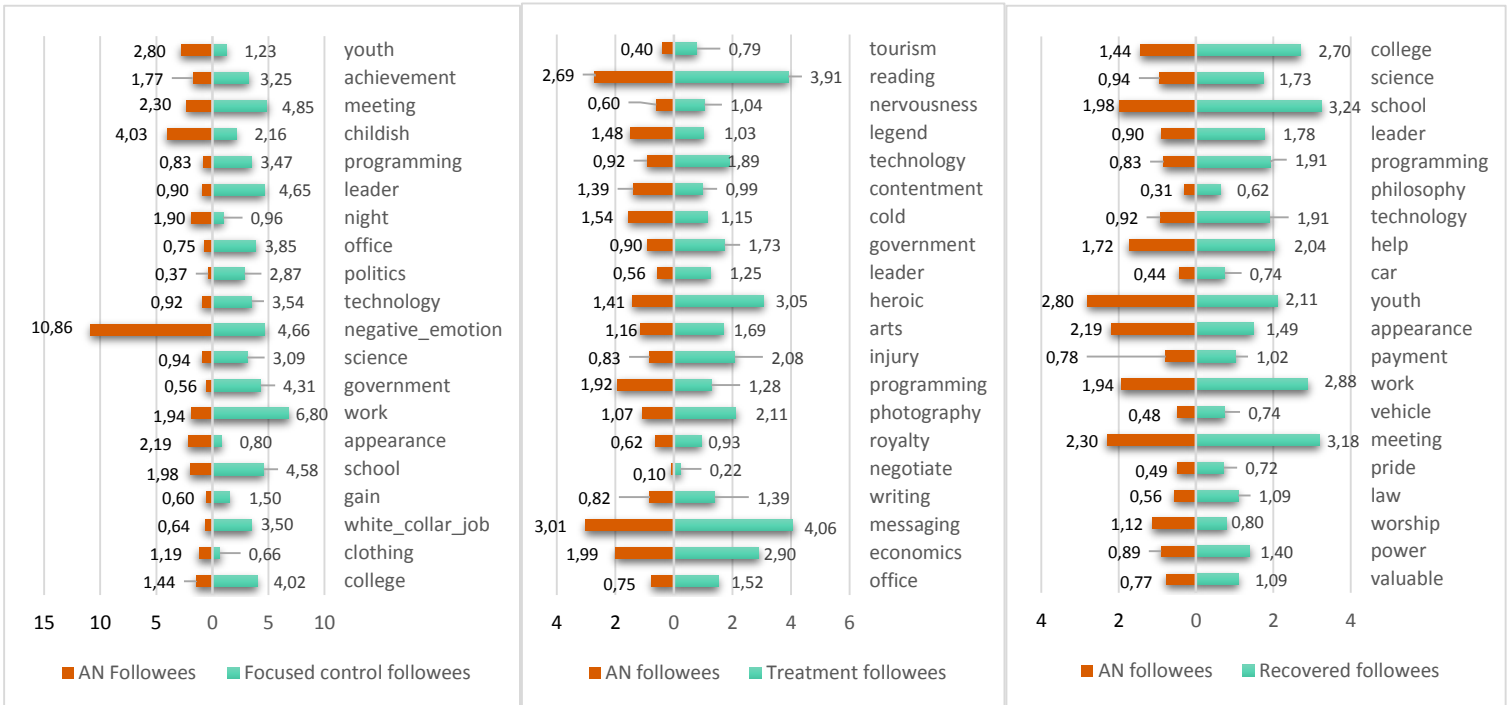

Figure 4.9. The top 20 topics with most significantly different values ($p<.05$) between the anorexia nervosa followees group and the focused control, recovered, and treatment followees groups. The median values for each feature are shown.

We observe that AN users and recovered users differ the least in their interests with their own followees. AN followees and focused control followees show the biggest difference in interests.

We also calculated the values for Spearman rank correlation coefficient based on the median values obtained for each topic in each group.

The pairs of groups were as follows: AN and AN followees (ρ=0.96), treatment and treatment followees (ρ=0.97), recovered and recovered followees (ρ=0.96), focused control and focused control followees (ρ=0.97), AN followees and treatment followees (ρ=0.93), AN followees and recovered followees (ρ=0.93), AN followees and focused control followees (ρ=0.69), treatment followees and focused control followees (ρ=0.86), and recovered followees and focused control followees (ρ=0.86).

From these results, we can say that for all the groups, their interests are highly similar to those of their followees; however, the interests of the treatment and recovered followees groups are more highly correlated to the focused control group followees than the AN followees group, indicating a change in interest through the evolution of the disorder.

c) Posting frequency aspects

We address the same features previously described for the characterization of suicidal ideation: the working week tweets count ratio, the weekend tweets count ratio, the median time between tweets, and the Sleep time tweets ratio (STTR).

The results of these behavioral aspects analyzed (Table 4.24) showed that AN users tweeted more on weekends compared with control groups.

In addition, the median time between tweets was lower for AN users (they tweeted more frequently) in comparison with random and focused control users.

We also observed that the tweeting ratio during sleeping periods was significantly higher for AN users than for the control groups. This might indicate some sleep alteration, which is a usual sign in EDs and other associated mental issues, such as depression.

Table 4.24. Comparative results between groups – Posting frequency aspects (***p<.001, **p<.01, *p<.05).

Feature	Median values					p-values and significance level (Mann-Whitney U)				
	AN	TREAT	RECOV	RAND CON	FOC CON	AN-TREAT	AN - RECOV	AN – RAND CON	AN – FOC CON	RAND – FOC CON
Working week tweets count ratio	0.73	0.73	0.75	0.75	0.75	.366	.152	.030 *	<.001 ***	.013 *
Weekend tweets count ratio	0.27	0.27	0.25	0.25	0.25	.366	.152	.030 *	<.001 ***	.013 *
Median time between tweets	625.25	701.00	1,187.50	4,063.75	1,088.00	.434	.149	<.001 ***	.005 **	<.001 ***
Sleep period tweeting ratio	0.05	0.05	0.04	0.04	0.03	.295	.066	<.001 ***	<.001 ***	.001 **
Normalized tweet count per year quarter - Dec-Feb	0.01	0.16	0.15	0.18	0.16	<.001 ***	.020 *	<.001 ***	<.001 ***	.013 *
Normalized tweet count per year quarter - Mar-May	0.01	0.23	0.21	0.23	0.21	.001 **	.106	<.001 ***	.014 *	.001 **
Normalized tweet count per year quarter -Jun-Ago	0.27	0.32	0.27	0.25	0.24	.387	.293	.100	.004 **	.063
Normalized tweet count per year quarter -Sept-Nov	0.36	0.24	0.27	0.28	0.31	.002 **	.200	.001 **	.323	.002 **
Median number of tweets created since the account creation	7,910.00	21,038.50	18,409.00	23,291.50	21,463.00	.004 **	.121	<.001 ***	<.001 ***	.494

Regarding the tweeting periods during the year, we see that between December and February (winter in Europe and summer in most countries of South America) users from the AN group tweeted less than users from all the other groups. However, we cannot match this finding to a clinical fact related to the seasons of the year, given the lack of information regarding the users' location.

d) Demographics

We analyzed the demographic characteristics (gender and age features) of the groups to verify whether these correspond to the actual incidence rates of AN [72]. These features are inferred, given the fact that Twitter does not publicly display the age and gender of users. We used the approach of Wang *et al.* [166] for demographic inference. This approach is based on a multimodal deep neural architecture for the joint classification of age, gender, and organizational status of social media users. Their model was trained using data in 32 languages, including Spanish. The method analyzes the description of a user and their profile picture.

We used the implementation of the method provided by the authors through a Python library named M3-Inference. The tool outputs scores for three gender categories—male, female, and organization—and four different age ranges.

Before using the detection tool on all the users, to increase its performance, and given the fact that the AN, treatment, and recovered users are not organizations, we defined that only if a user had a score over 0.70, for the organization class, and if this value was higher than the scores for males and females, then this label would be assigned; otherwise, the maximum value among the male and female scores was considered. In addition, if the organization label was assigned to a user, we automatically assigned a specific age group (classified as an organization) for all the users classified as organizations. We evaluated the performance of this approach on a group of manually labeled users based on their translated descriptions, where we considered up to 50 users per group.

We obtained a macro average accuracy of 0.84 for all the gender groups of all the classes and a macro average accuracy of 0.80 for all the age groups of all the classes.

We obtained the percentages of users corresponding to demographic categories for each group (Figure 4.10). Most of the AN, treatment, and recovered users were young women. These results are compatible with the statistics that mention that the incidence rates for AN are the highest for women aged 15-19 [72]. We also observe that a considerable number of users in the focused control group represent organizations. When comparing the ratios of users belonging to each gender per group (Table 4.25), we see differences between the AN and control groups due to the number of female users. We also observed differences between the focused control and random control groups, as there were fewer organizations in the random control group. Regarding age (Table 4.26), we observe that AN users differ from the control groups because of the large number of AN users aged ≤18 years. We also find differences between the AN and recovered groups, as recovered users are normally older than AN users. This is consistent with the fact that a full recovery process often takes years, and therefore, users get older as the recovery stages are reached.

Table 4.25. Comparative results between groups – Gender groups (***p<.001, **p<.01, *p<.05).

Gender	p-values and significance level (Proportions Z-test)				
	AN - TREAT	**AN - RECOV**	**AN – RAND CON**	**AN – FOC CON**	**RAND – FOC CON**
Male	.150	.141	<.001***	.436	<.001***
Female	.150	.141	.044*	<.001***	.501
Organization	--------	---------	<.001***	<.001***	<.001***

Table 4.26. Comparative results between groups – Age groups (***p<.001, **p<.01, *p<.05).

Age group	p-values and significance level (Proportions Z-test)				
	AN - TREAT	**AN - RECOV**	**AN – RAND CON**	**AN – FOC CON**	**RAND – FOC CON**
≤18	.321	.010*	.002**	<.001***	.783
19-29	.274	.042*	.313	<.001***	.086
30-39	.801	.062	<.001***	.009**	.020*
≥40	.495	.585	<.001***	<.001***	.003**
Classified as organization	--------	---------	<.001***	<.001***	<.001***

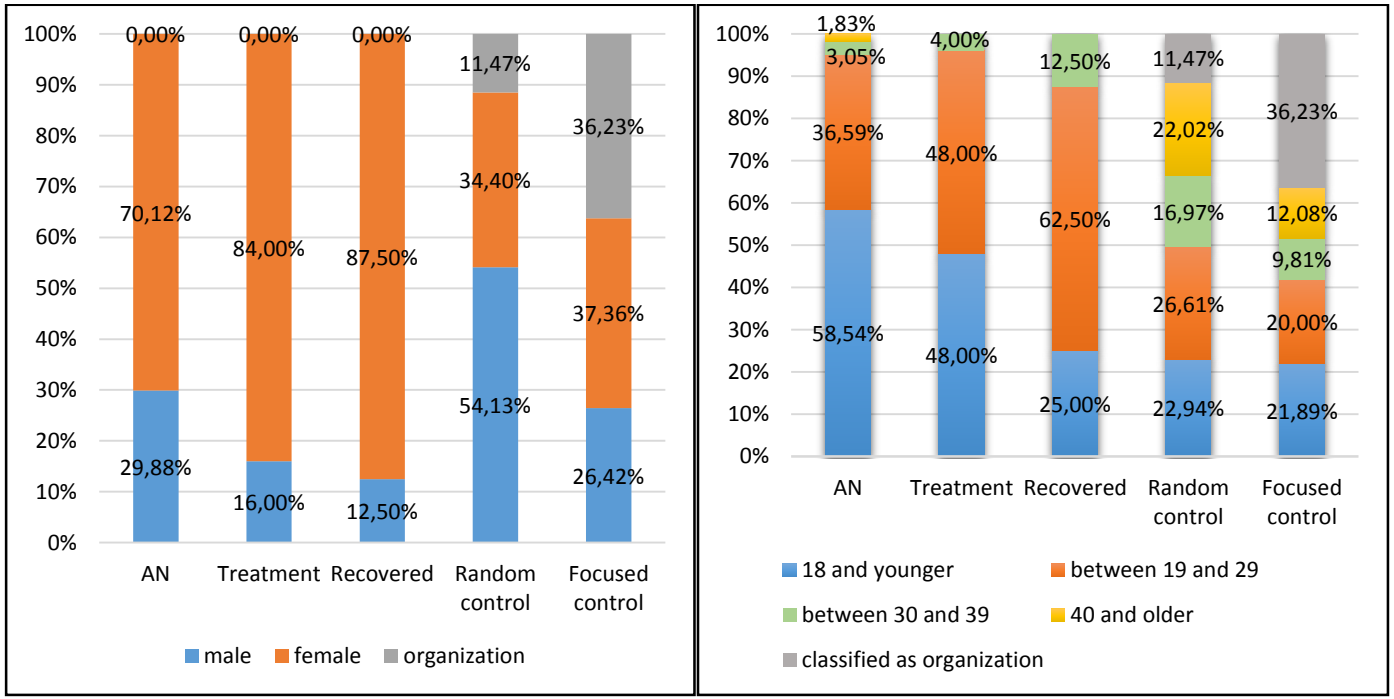

Figure 4.10. Composition of the anorexia nervosa, treatment, recovered, and control user groups according to gender and age. Each age and gender subgroup is represented by a color.

e) Visual aspects

We describe the use of features extracted from the profile pictures of users and from the images of posts shared by users. For the first case, we use pre-trained models provided by external sources; for the second case, we analyze the output of a model trained on our data set and apply it to the pictures shared by a set of validation users from our groups of interest. We explore visual aspects given that there are physical traits that characterize AN [12].

- Profile picture

We analyzed 32 features extracted from the pictures of users. We explored the technical features and the detection of emotions and objects.

As part of the technical features analyzed, we checked if an image is grayscale, if it is lighter, if it has text, and if it has faces on it. We also analyzed the existence of objects in the pictures. These features are defined through the use of Python libraries such as imageio [148] for verifying the brightness of an image, PIL [123] for verifying if an image is grayscale, and the General Recognition AI API from Chooch AI [4], which after taking an image as an input, it outputs the names of elements recognized in the picture, such as texts, clothes, faces, animals, and specific objects.

We also detected emotions expressed on the pictures using the Algorithmia facial emotion recognition API, which implements CNN models [88]. The models included in the previous APIs were already trained, so we only ran them over the profile pictures in our data set, and no rights over a further use of these images were granted to the tools' owners. In addition, none of the images were observed by any human annotator, and only the extracted features were stored.

To analyze the existence of objects in the image and represent a user through these features, we considered a Boolean BoW model. In this model, the names of the objects found were the terms considered, and the value for an object was assigned as 1 if it was found on the picture and 0 if it was not found. Given the sparsity of the model, we only considered objects that appeared in at least five images, which led us to retain 20 features of this type. Regarding the emotions' features, we assigned to a user (if there are faces in the picture) the emotion with the highest predictive score, and later, to define a

score for a group, we considered the ratio of users assigned to a given emotion. The same approach was considered for the technical features and objects detected.

There are significant differences between the groups' profile pictures. Regarding the technical aspects (Table 4.27), focused control users are likely to be distinguished from AN users because of the presence of text in their profile pictures. This also applies to random control users, who tend to use text as well but in a lower ratio than focused control users. These findings can be explained by the use of logos in the profile pictures of the accounts of organizations.

In addition, AN users' pictures are significantly darker than focused control users' pictures. In terms of the emotions detected (Table 4.28), treatment users expressed more neutral emotions. Sadness characterizes AN users, which are the only ones with a ratio of users showing such expressions.

On the objects detected (Table 4.29), there were significant differences in the existence of clothing elements between the AN and control groups, along with the appearance of hands, shorts, and accessories, which might suggest that more full-body pictures are shared by AN users, which might consequently imply a higher interest in their appearance.

There is a high ratio of posters on the control users' profiles, which validates our prior assumption about the representation of organizations. Few men were identified on pictures of users of the AN group, whereas women were identified on more than half of the AN profile's pictures.

Table 4.27. Comparative results between groups – profile picture: technical aspects (***p<.001, **p<.01, *p<.05).

Feature	Users ratio					p-values and significance level (Proportions Z-test)				
	AN	TRE	REC	RAND CON	FOC CON	AN - TRE	AN - RECOV	AN – RAND CON	AN – FOC CON	RAND – FOC CON
Is gray scale	0.094	0.125	0.167	0.0521	0.0310	.778	.569	.308	.064	.425
Is lighter	0.500	0.625	0.500	0.4167	0.6589	.505	1.000	.299	.033 *	<.001 ***
Has text	0.063	0.125	0.000	0.2188	0.3411	.512	.528	.008 **	<.001 ***	.045 *
Has faces	0.281	0.500	0.333	0.2500	0.2636	.205	.787	.660	.794	.818

Table 4.28. Comparative results between groups – profile picture: Emotions detected (***p<.001, **p<.01, *p<.05).

Feature	Users ratio					p-values and significance level (Proportions Z-test)				
	AN	TREAT	RECOV	RAND CON	FOC CON	AN – TREAT	AN – RECOV	AN – RAND CON	AN – FOC CON	RAND – FOC CON
Neutral	0.000	0.375	0.000	0.0521	0.0853	<.001 ***	--------	.064	.016 *	.338
Sad	0.094	0.000	0.000	0.000	0.000	.366	.433	.002 **	<.001 ***	--------
Fear	0.047	0.000	0.167	0.0313	0.0155	.532	.227	.610	.196	.428
Surprise	0.016	0.000	0.000	0.000	0.0078	.722	.758	.219	.611	.387
Angry	0.094	0.125	0.000	0.0625	0.0698	.778	.433	.462	.558	.829
Happy	0.031	0.000	0.167	0.0833	0.0930	.612	.117	.182	.119	.801
Disgust	0.000	0.000	0.000	0.0313	0.0155	--------	--------	.153	.317	.428

Table 4.29. Comparative results between groups – Profile pictures: objects detected (***p<.001, **p<.01, *p<.05).

Feature	Users ratio					p-values and significance level (Proportions Z-test)				
	AN	TREAT	RECOV	RAND CON	FOC CON	AN – TREAT	AN – RECOV	AN – RAND CON	AN – FOC CON	RAND – FOC CON
Poster	0.000	0.000	0.000	0.0938	0.2403	--------	--------	.012*	<.001***	.004**
Clothing	0.750	1.000	0.500	0.5313	0.3333	.109	.188	.005**	<.001***	.003**
Person	0.281	0.125	0.000	0.3229	0.1240	.344	.132	.575	.007**	<.001***
Man	0.047	0.000	0.000	0.3229	0.1550	.532	.588	<.001***	.029*	.003**
Dress	0.031	0.000	0.167	0.0104	0.0233	.612	.117	.341	.742	.471
Boy	0.016	0.000	0.000	0.0208	0.0078	.722	.758	.812	.611	.397
Tree	0.016	0.000	0.167	0.0521	0.0310	.722	.034*	.234	.527	.425
Human hand	0.063	0.000	0.000	0.000	0.0078	.467	.528	.013*	.024*	.387
Fashion accessory	0.078	0.000	0.000	0.0208	0.0155	.412	.477	.083	.028*	.765
Flower	0.031	.0.000	0.000	0.0104	0.0388	.612	.660	.341	.793	.192
Glasses	0.000	0.125	0.000	0.0521	0.0310	.004**	--------	.064	.155	.425
Animal	0.000	0.000	0.000	0.0208	0.0233	--------	--------	.245	.219	.903
Shorts	0.063	0.125	0.000	0.000	0.000	.512	.528	.013*	.004**	--------
Jeans	0.031	0.000	0.000	0.0208	0.0078	.612	.660	.679	.214	.397
Human eye	0.063	0.000	0.000	0.0104	0.0078	.467	.528	.064	.024*	.833
Cat	0.000	0.125	0.000	0.0104	0.0233	.004**	--------	.413	.219	.471
Footwear	0.094	0.000	0.000	0.0521	0.0155	.366	.433	.308	.010*	.118
Human nose	0.016	0.000	0.000	0.0208	0.000	.722	.758	.812	.155	.100
Girl	0.188	0.250	0.167	0.0521	0.0310	.674	.900	.006**	<.001***	.425
Woman	0.578	0.750	0.667	0.1771	0.2248	.350	.674	<.001***	<.001***	.380

- Pictures shared

We explored the pictures shared by users through their individual posts to detect AN-related images. For this purpose, we analyzed the output of two models trained on the images shared by users from three of our groups of interest: AN, focused control, and random control. Two binary classification models were trained by members of the Computer Vision Center of the Unversitat Autónoma de Barcelona: (1) an AN versus focused control image detection model and (2) an AN versus random control image detection model [138]. The resulting model was applied to all the images of a set of users kept for validation purposes, which were not considered in the training process. The output of each model was a score for each class to predict. For a single user, the score corresponding to this feature is given by the average score obtained by the classifier for the AN class over all the user's images. For the first model (AN vs focused control), 278 users were considered for training and 130 for validation purposes. For the second model (AN vs random control), 240 users were considered for training and 106 for validation purposes.

The results showed that there were highly significant differences between the AN and control groups ($p<.001$ for both comparisons). The median of the aggregated scores of the first classifier for the AN class (AN vs focused control) for a set of 130 validation users was 0.73, whereas the median value for focused control users was 0.36. This means that a higher number of pictures related to AN were found on the posts of AN users. When analyzing the median of the aggregated scores of the second classifier (AN vs random control), on a set of 106 validation users, the median value for AN users was 0.78 and for random control users was 0.54. We observed lower aggregated scores for both control cases, meaning that these users share fewer AN-related pictures. These results show that the pictures can be informative for the detection of users with AN.

4.4.3 Insights of users at the contemplation stage

In this section we perform a deeper analysis of users at the contemplation stage [124], given that according to the trans theoretical model, these correspond to people that are

considering changing unhealthy habits, and thus seek help and eventually treatment. In this context we 1) explore the main terms and topics of interest addressed by people with anorexia nervosa at the contemplation stage. 2) Through a survey, we ask volunteers that have gone through the contemplation stage about their interests during that stage, and compare how the interests extracted from social media data differ from those provided by volunteers. 3) As it is our hypothesis that social platforms reinforce the suggestion of harmful accounts to users with anorexia, we take Twitter as a use case and measure the percentage of harmful and harmless accounts suggested by Twitter's recommender system to AN users.

a) Analysis of topics of interest

A total of 22 participants with AN, at advanced stages of treatment were reached through a specialized recovery center. In order to obtain the topics of interest of participants we first analyzed the topics of interest of contemplation users in Twitter. To do so, using Dataset 3, two clinicians were asked to label the set of 171 AN users as either precontemplation or contemplation cases. We considered as contemplation cases the ones for which both of the annotators agreed (56). Then we obtained the interests of the contemplation users. The process followed to obtain the topics of interest of these users was the same as the one used in Section 4.4.2 for the analysis of the topics of interest of users of the other stages of AN. For this case, instead of using Empath as a tool, the resource used to extract the topics was Dandelion's entity extraction API[6], which given a text, it extracts key n-grams and returns Wikipedia's and DBLP's categories to which a term or n-gram belongs to. Compared to other recommendation methods based on content, which often make exclusive use of n-grams (terms), Dandelion provides us with semantic topic categories that are more general, yet specific enough to recognize and categorize elements. Instances are brands and artists' names that are recognized and placed in dedicated categories.

Later, the top 200 topics of interest of all the users were obtained and manually classified into broader topic categories and subcategories. These categories were defined in order to design a survey for the participants.

[6] *Dandelion Entity extraction API - https://dandelion.eu/docs/api/datatxt/nex/v1/*

Participants were asked to assign a level of relevance to each subcategory according to their interests during the contemplation stage. Interest levels were between [0,5], where 0 meant a lack of interest and 5 implied a very high level of interest (Table 4.30).

Table 4.30. Categories and subcategories defined from the topics of interest of contemplation users.

Main categories	Subcategories
Technology	• Applications, social networks and social media social media • Technological devices • Video games
Health	• Nutrition • Physical wellbeing • Mental wellbeing
Lifestyle and personal beliefs	• Interpersonal relationships • Activism • Religion and spirituality • Economy • Politics and justice
Science	• Philosophy • Sociology • Biology • Cosmology • Chemistry • Physics and Mathematics
Hobbies and entertainment	• Sports • Movies and TV • Music • Literature
Other interests	• Current news • Languages • Cultures of the world • Others

Participants were also asked to freely specify their particular interests during the contemplation phase, within each of the sub categories mentioned, for instance, video games. If participants assigned a certain level of interest to the subcategory, they would specify keywords of interest for them related to the topic (*e.g.,* Mario Kart, Nintendo, etc.).

The volunteers were also asked to describe the topics of interest that they think were mostly related to AN, and if they thought that the suggestions made by social platforms were harmful or not for them.

From this outcome we proceeded to extract the topics of interest of the participants. We obtained their topics based on the specific keywords they used to describe their interests for each given category.

Instead of assigning a score to each topic using the frequency of the keywords related to it (as they were only mentioned once), we assigned the level of importance assigned by the participant to the subcategory evaluated (see Table 4.30). As an example, if a participant, within the video games subcategory had assigned a level of interest of 4 to the subcategory, and reported “play station” as a keyword describing their particular interest, the topic detection tool would automatically assign this keyword to its own categories like game console, video games, Sony consoles, etc., and then we would assign to all of these categories a score of 4.

Later, we represented all the interests of all participants through a bag of words/topics model, where the scores assigned to each (participant, topic) were scaled between 0 and 1 based on the max and min scores for all the topics, according to each given participant. This way, we obtained a vector of scored topics representing each participant.

With the data collected from the survey applied to participants, we analyzed the results regarding the scores given by participants to each topic of interest (Table 4.30), and obtained the topics that are relevant for them by aggregating the results obtained by each participant and summarizing our findings in a box plot.

In Figure 4.11 we report on the level of importance assigned to each subcategory predefined in Table 4.30. We observe in the boxplot that the main topics of interest are nutrition, music, physical wellbeing, apps, mental wellbeing and interpersonal relationships.

Regarding the topics of interest mostly related to AN for participants, we analyzed the frequencies of terms used in the answers of users and represented these terms and their importance in a word cloud, where the terms or bigrams most used are displayed in major size.

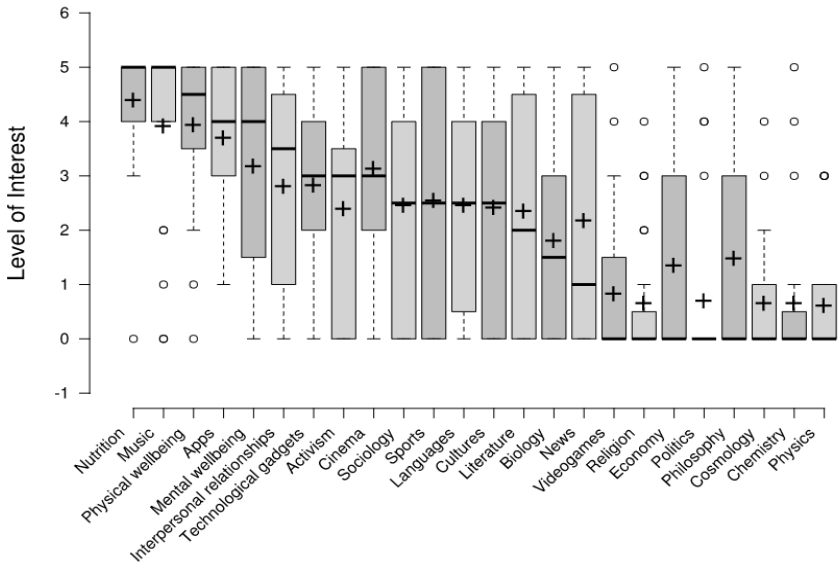

Figure 4.11. Levels of interest of participants in the predefined topics addressed in the survey.

A total of 77.27% (17/22) of the participants surveyed thought that the content suggested by social platforms was harmful for them.

Also, the topics related to AN that they were most interested in are described in Figure 4.12, where we can see the top terms (translated from Spanish to English) used to describe their topics of interest. We can see in the word cloud that terms such as exercise (freq=9), diet (freq=9), lose weight (freq=9), food (freq= 6) and physical (freq=4) are the most mentioned.

Figure 4.12. Word cloud obtained from the descriptions provided by survey participants (N=22) of their topics of interest related to AN.

Finally, we established a comparative analysis of the topics automatically extracted from participants and those extracted for the data of Twitter users. From the topics of Twitter users, and those of the participants, we show the top 10 topics of interest. Topics were ranked based on their frequencies. Following the same approach we also obtained the top 10 terms most used by each group. As it can be seen in Table 4.31, the top 10 topics of interest of users and survey participants are quite similar and can easily be related to anorexia nervosa. Moreover, four topics can be found in both groups. Regarding the terms mostly used by participants and users we can see again that most of them are related to anorexia nervosa.

Table 4.31. Most addressed topics and most terms used by contemplation Twitter users and survey participants.

Rank	Participants' topics	Twitter users' topics	Participants' terms	Twitter users' terms
1	**Nutrition**	**Software**	Eating disorders (*tca*)	Fasting (*ayuno*)
2	**Social Networks**	**Social Networks**	Diets (*dieta*)	Fat (*gorda*)
3	Medical terms	Twitter	Instagram	**calories (*calorías*)**
4	Culture sociology	**Food**	Weight (*peso*)	I ate (*comí*)
5	**Software**	Vegetarian food	**calories (*calorías*)**	Day (*día*)
6	**Food**	**Nutrition**	Lose weight (*adelgazar*)	Eating (*comiendo*)
7	Images storage	Dairy food	Food (*alimentos*)	Pretty (*linda*)
8	Diets	Fruits	Exercise (*ejercicios*)	Hours (*horas*)
9	Energy units	Culinary ingredients	Series	Say (*decir*)
10	Measurement units	Internet	Self-image (*imagen*)	Horrible

b) Analysis of the type of users recommended by Twitter to users with anorexia

We analyzed Twitter's recommendation method considering it as a black box and analyzing the recommendations given by the platform. In this sense we consider, three types of accounts: 1) harmful accounts are those that can negatively influence the behavior of users with anorexia, here we can find accounts that promote diets and excessive exercising, accounts that express concerns about body image and promote unhealthy eating habits, and specially pro-ED accounts, among others. 2) Pro-recovery accounts correspond to specialized recovery centers, educational psychologists, foundations and people that can offer support and information towards recovery from eating disorders. Finally, 3) neutral accounts are those that do not promote harmful nor pro-recovery content. We consider then that harmless accounts are the union of neutral and pro-recovery accounts.

The steps followed to measure the percentage of accounts of each type suggested by Twitter to users with anorexia were: 1) among 50 twitter AN-Contemplation labeled accounts, we have labeled the followees (50 per each account) of the accounts as either harmful, neutral, or pro-recovery accounts (2,500 users in total). 2) We obtained the average number of accounts of each type followed. Then, 3) we have also created 20 Twitter accounts to reproduce the process of following accounts by ED users, and evaluated the types of accounts suggested by Twitter to follow. For each of the 20 accounts, we followed 50 accounts. From these, a percentage corresponded to harmless accounts and another percentage corresponded to harmful accounts (based on the ratios obtained from step 2). For the harmless accounts, users were followed based on the initial suggestions given by Twitter once an account is created. Regarding the harmful accounts, with the keywords: *edtwt, proana, promia* as search terms, we searched for harmful user's accounts and randomly followed the corresponding percentage of accounts suggested according to the search terms. Later, based on these 50 accounts followed, we labeled the top 50 accounts suggested by Twitter in their "who to follow" section as either harmful, pro-recovery or neutral (1,000 users labeled).

Finally, after evaluating the ratio of harmful, neutral and pro-recovery users followed by contemplation users in Twitter, we obtained the results described in Table 4.32. There are only a few harmless users (18.52%) and there are no pro-recovery accounts among them. Also, on average 73.70% of the accounts suggested by the platform to people with AN are likely to be harmful.

Table 4.32. Types of users followed by Twitter's AN contemplation users.

Label	followees (mean)	followees (median)	followees (%)	users suggested (mean)	users suggested (median)	users suggested (%)
Harmful	40.74	41.5	81.48% (40.74/50)	36.85	36.5	73.70% (36.85/50)
Neutral	9.26	8.5	18.52% (9.26/50)	12.65	12.5	25.30% (12.65/50)
Pro-recovery	0	0	0% (0/50)	0.5	0	1% (0.5/50)

4.5 Discussion

In Chapter 4 we have addressed our first research question dedicated to the characterization of mental disorders at a post and at a user level using data mainly from two different social platforms: Reddit and Twitter, with data collected in English and Spanish.

We have obtained several insights regarding the use cases studied through the analysis of multiple perspectives that mainly imply the extraction and inference of multimodal, behavioral and demographic elements.

One of the first aims of this work was to determine lexical features characterizing each of the use cases studied. Results show that there are many elements that distinguish writings of control users (CON) from those of people with certain mental disorders and substance abuse conditions (MEN), while there are fewer elements that distinguish the conditions analyzed (SUI, DEP, ED and ALC) from each other. Notably, terms related to emotions and feelings are expressed more in writings of the MEN class, while words concerning topics such as work, money, and home are more frequent in the CON class writings. We can also observe that, as expected, the categories

that imply risk factors for mental disorders such as self-harm, suicidal ideation references, self-hatred, substances abuse, lack of social support, bullying or other types of abuse, obtained higher scores for the MEN group, providing evidence of the fact that the aspects that are considered on screening processes [106] can also be identified on social media posts.

We see that some of the categories addressed can characterize exclusively certain conditions such as the highly significant expression of negative emotions by the SUI class; the references to caloric restrictions, body image, laxatives, and body weight of the ED class; the references to antidepressants of the DEP class; and the reference to topics related to leisure activities, which often involve drinking, for the ALC class.

In order to distinguish differences in the vocabulary used by each group, we have also proposed a process to identify specific terms (unigrams and bigrams) that characterize exclusively a given group. An instance can be the usage of the bigram *alcoholic anonymous,* which is likely to characterize the ALC group. This approach is important for the characterization of the conditions studied, and is also useful to improve the representation of text using word embeddings for their usage adapted to a task dedicated to the detection of mental disorders in social platforms. This will be proved in Chapter 5.

Having performed a deep analysis of suicidal ideation, we have identified that behavioral and psychological features are relevant to distinguish control from risk cases. Discussion topics such as money and work, were mostly addressed by generic control users, whereas the members of the suicidal ideation risk group use terms more related to health and biological aspects. As for the MEN group in our Reddit analysis of posts in English, the use of self-references was higher in suicide risk cases compared to both control groups (usage of singular first person personal pronouns and verbs conjugated in singular first person).

Behavioral aspects regarding tweets statistics, posting patterns, and the interactions between users were relevant too. At a 95% CI, when comparing the suicidal ideation risk group and the focused control group, the number of friends ($p=.04$) and median tweet length ($p=.04$) were significantly different. The median number of friends for focused control users

(median 578.5) was higher than that for users at risk (median 372.0). Similarly, the median tweet length was higher for focused control users, with 16 words against 13 words of suicidal ideation risk users. Also the Sleep time tweets ratio (STTR), as for people with AN, was higher compared with control cases.

Finally, the analysis of images suggested that they could be relevant to distinguish risk cases from generic control cases, while for the case of anorexia nervosa the use of images was relevant to distinguish AN cases from both: generic (random) and focused control cases.

Regarding anorexia nervosa, we introduced the first analysis with social media data of the different stages towards recovery using the TTM. We found multiple elements that characterize and distinguish users with AN at the early, treatment, and full recovery stages.

AN users tweet more frequently at night and on weekends in comparison to focused control users. These results are consistent with clinical findings that suggest that patients with AN often report poor sleep quality and reduced sleep time [109]. In addition, the image results indicate that the analysis of visual elements is relevant for the detection of AN cases and focused control cases. In particular, our findings showed that the features extracted from the content generated by users are the most relevant for characterizing AN users, especially those related to linguistic and psychological factors, including terms that describe risk factors and the signs and symptoms of AN (*i.e.*, anorexia-related vocabulary).

We also determined the linguistic attributes that characterize Spanish-speaking users with AN and found that similar to related work on English texts [6], and to our findings with other disorders, the high use of first-person singular pronouns and verbs conjugated with these pronouns distinguishes AN users from control users. It was relevant to find that it also characterizes AN users when compared to recovered cases ($p<.001$), which make more use of first person plural pronouns. We have also observed that the AN group is characterized by a significantly lower use of articles and higher use of impersonal pronouns than control users ($p<.001$), which can be a particular characteristic of the language usage (Spanish).

Our findings also reinforced the relevance of textual elements through the development of a deep learning model for the detection of AN-related tweets. We explored the change in the ratio of posts related to AN tweeted by users at each stage and found that users at the early stages of AN posted more AN-related tweets, which also happened for the suicidal ideation case with the scores obtained by the tweets of users at risk for the SPVC. In addition, highly significant differences among AN cases and recovered cases, and control cases ($p<.001$) were observed. There were also very significant differences between the AN and treatment groups ($p=.004$) in terms of this feature. It implies that the proportion of tweets related to AN significantly changes depending on the recovery stage, indicating the progress in the recovery process of social media users with AN.

Among the relational factors explored, we found that AN and focused control groups could be identified by analyzing the structure of their social network (clustering approach). Among focused control users, there were several organizations and specialists for the treatment and prevention of EDs. The high polarization noted among the AN and focused control communities reinforces the findings of previous studies conducted on networks of English speakers [163,164], which reported limited interactions between ED and pro-recovery communities. From a psychological perspective, these findings can be explained by the elements that characterize people at the precontemplation stage according to the trans theoretical model of health behavior change, where people are in denial of their unhealthy conditions and tend to feel supported by their equals (pro-ED community members) [69,122], resulting in a rejection of pro-recovery content.

Regarding the topics of interest of users and their followees, we found that the interests of AN users and their followees were highly correlated ($\rho=0.96$). We also observed a higher correlation between the treatment followees and focused control followee groups ($\rho=0.86$) and the recovered followees versus focused control followees groups ($\rho=0.86$) in comparison with the AN followees versus focused control followees groups ($\rho=0.69$). These results show that more interests are shared with focused control users as the recovery process advances. In this sense, our findings suggest that

there is a willingness of people with AN at the contemplation stage to seek assistance, as we have seen that eventually, as the treatment and recovery stages are reached, the polarization is reduced.

Finally, when we explore the existence of differences among our control groups (random and focused control users), we observed that the focused control group had three times more organizations' accounts than the random control group. We also noticed that the main differences among these groups were found in linguistic attributes, especially for focused control users that were characterized by the use of a reduced number of swearing terms and more anorexia-related vocabulary terms. These findings complement our prior assumptions that focused control users were mostly organizations, specialists, and clinicians, corresponding to a pro-recovery community [33,163,164].

Given our findings, another relevant aspect to discuss is the indirect and unintended role of social platforms in the promotion of harmful content among users with and without EDs. A previous study [78] reported significant decreases in caloric intake among people exposed to pro-ED websites between pre-exposure and post-exposure. As recommender systems are designed, users with AN are likely to be recommended to follow users similar to them (other AN users), and this indirectly contributes to the reinforcement of unhealthy habits.

We have performed a further analysis of the contact recommendation provided by social platforms to users with anorexia nervosa through a survey dedicated to people in treatment from anorexia and through the reproduction of the activity of people with anorexia in Twitter. A total of 77.27% (17/22) of the survey participants thought that the content suggested by social platforms was harmful for them. While our analysis of Twitter data found that on average 73.70% of the accounts suggested by the platform to AN users are likely to be harmful.

The findings of this analysis encourage the creation of detection and risk aware recommendation tools to assist people with mental disorders, and these are the aspects that we explore in the following chapters of this thesis.

Chapter 5

PREDICTIVE MODELS

5.1 Introduction

In this chapter we describe several predictive models for mental health state assessment. We define and evaluate specific models for the detection of anorexia, depression and suicidal ideation.

We take into account cases of users with mental conditions and control cases to build our predictive models. We also design models dedicated to identify cases of a given mental disorder over multiple types of disorders.

The models evaluated make use of several multimodal and behavioral features that measure elements from images, the social network of users, their posting frequencies, and vocabulary usage.

We use methods to represent text elements, in particular we focus on the usage of word embeddings for which we propose an approach to enhance the usage of these representations taking into account the predictive tasks that they will be used for.

Finally, we address work dedicated to the development of early risk detection systems and provide insights regarding gender biases assessment in predictive models dedicated to the detection of anorexia.

5.2 Suicidal ideation assessment

5.2.1 Introduction

In this section we address suicidal ideation detection [126]. We evaluated statistical and deep learning approaches to handle multimodal data for the detection of users with signs of suicidal ideation (suicidal ideation risk group). Our methods were evaluated over Dataset 2 – Suicidal ideation, which consisted in a dataset of 252 users annotated by clinicians.

As part of this work we: 1) generated models that explore the impact of not just relational and behavioral factors but also elements identified by specialists during consultations, which have been mapped to social networks. 2) We developed image-based predictive models to detect suicidal ideation. 3) We integrated the previous elements into a method that combines multimodal data to build predictive models that address the detection of mental health issues; and 4) we refined the evaluation process of predictive models for mental health issues by considering 2 different types of control groups within the social media context: users with posts that might not use terms related to mental conditions (generic control cases) and users who make use of such terms (focused control group).

5.2.2 Models description

We propose a method that given the profile of a user: (1) it uses a text-based model, described previously as the SPVC (Section 3.4.1), which selects a subset of relevant tweets related to suicidal ideation. The set of tweets for which the SPVC provides a score over a given threshold is retained in the short profile version (SPV) itself; (2) mostly from the outputted SPV, it extracts a set of relational, textual, behavioral, lexical, statistical, suicidal ideation–related, and image-based features from the content and metadata of the tweets; and (3) it builds and evaluates different predictive models resulting from the combination of these features.

We make use of all the features previously described in Section 4.3.2. These features are organized into 3 different groups: 1) BoW or n-grams and word embeddings as a representation of textual features; 2) a set of behavioral features known as social networks and psychological (SNPSY)

features containing a group of relational, posting frequency-based, lexical, sentiment analysis, and statistical features, in addition to a set of features that attempt to map to the social media context certain signs and symptoms, which are usually considered by clinicians at the time of screening; and 3) an image-based score.

a) Classification tasks

As we wanted to evaluate the change in the performance of models that use 2 different types of control groups: focused control and generic control, we created experiments for comparing 1) users at risk versus focused control users (task 1) and 2) users at risk versus generic control users (task 2). These were selected as our 2 supervised predictive tasks.

b) Baselines

We defined as baselines 2 models exclusively based on generic text representations. These models were generated using the features extracted (Section 4.3.2) and representations from the users: 1) full profile and 2) from their SPV. The first one is a BoW model trained with {1-5}-grams, and the second one consists of a deep learning model defined by a CNN architecture that has been proven to be successful for text classification [82]. It also has been used in a similar task that addresses suicide risk assessment on Reddit users [146]. For this model we adopted the approach of Shing *et al*. [146] to define our user-level instances.

Given a user represented by a set of sequential posts, we concatenated all these posts and represented each post as a concatenation of words, where each word is represented by a vector (word embedding). As in the study by Coppersmith *et al.* [39], we used a set of word embeddings previously learned on Twitter [43] to define the starting weights for our embedding layer and performed further fine-tuning to learn over the training set and adapt the representations to the task domain.

We considered the 2 models previously described as state-of-the-art approaches for the creation of generic and exclusively text-based models for the task, as it is one of the purposes of our work to analyze the contribution of the additional feature types defined. We therefore defined 4 baseline models. Baselines 1 and 3 correspond to the BoW model generated over the full profile tweets sample and the SPV, respectively. Baselines 2 and 4 correspond to the deep

learning model built over the same data samples (full profile and SPV).

c) Classifiers

With the intention of evaluating the individual contribution of the types of features defined, along with their combinations toward a classification/detection task, we used 4 types of classification algorithms and a deep learning model: random forest, multilayer perceptron, logistic regression, and support vector machines as classifiers.

For each feature combination approach, models were built for all these classifiers using the Scikit-learn [32] library's implementation, with a grid search for the best parameters. We also used a CNN architecture fed by embedding models.

d) Approaches for combining features

We evaluated several ways of combining our 3 main feature types defined: generic text-based features, SNPSY features, and the image-based feature (image user score).

As can be seen in Table 5.1, we first generated individual models using exclusively all the features corresponding to the BoW model, the embedding model, and the SNPSY model, with features mainly obtained from the users' SPV.

Afterward, we explored the combination of our different feature types using the BoW model to represent text-based features.

Our first approach involves combining the BoW features with the SNPSY features. In this case, given the large number of BoW features and their sparsity, we opted to use the BoW model–predicted probabilities as values for a single feature, denoted as the BoW outputted feature, to be added to the SNPSY set of features. This is described in Table 5.1 as the *(BoW + SNPSY) model*.

Subsequently, we evaluated the combination of the BoW features with the image feature. For this case, we simply added to the BoW set of features the image user score as another attribute; this combination is described by the *(Images + BoW) model*.

Afterward, to combine the SNPSY features with the image feature, we used the image user score as a new feature in addition to the SNPSY feature set, which is the *(Images + SNPSY) model*.

Table 5.1. Models and features.

Model	Features	Number of features	
		Task 1	Task 2
BoW model	Bow features generated with the TF-IDF vectorizer with {1-5}-grams features.	24,645	24,336
Embeddings model	Word embeddings representations as input for a text-based convolutional neural network model	200	200
SNPSY model	SNPSY features = posting frequency + relational + tweets statistics + lexicons + suicide risk factors vocabulary + sentiment analysis features.	112	112
(BoW + SNPSY) model	BoW outputted feature + SNPSY features	24,757	24,448
(Images + BoW) model	Images user score + BoW features	24,646	24,337
(Images + SNPSY) model	Images user score + SNPSY features	113	113
(Images + BoW + SNPSY) model 1	Ensemble model = images user score + BoW outputted feature + SNPSY outputted feature	24,758	24,449
(Images + BoW + SNPSY) model 2	SNPSY features + images user score + BoW outputted feature	114	114
Selected features' model 1	Selected features from all the feature types with $p<.05$	5,807	14,882
Selected features' model 2	Selected features from all the feature types with $p<.001$	522	3,250

Finally, to combine the 3 feature types, we defined 2 approaches. The first approach is an ensemble model where we consider the outputs (predicted probability scores) of the BoW model (BoW outputted feature) and SNPSY model (SNPSY outputted feature) along with the image user score. This approach corresponds to the *(Images + BoW + SNPSY) model 1* with 3 attributes based on the combination of the 3 independent models with all their features.

The second approach consists of using all the features of the SNPSY type as attributes in addition to the BoW output feature and the image user score, which lead to the definition of *(Images + BoW + SNPSY) model 2*. It is necessary to recall that the predicted probability scores from the BoW and SNPSY individual models that were used for some of the feature combination approaches at the training stage correspond to the outputs of the classifiers on the test folds during the cross-validation process executed on the training set. This was done to avoid overfitting.

In addition to the combination approaches described, we created 2 other models over which we performed a feature selection procedure over all the feature types. We chose the features with statistically significant differences among the suicide and control groups to evaluate their contribution exclusively to a predictive model. We presented 2 models with features selected based on the p-values obtained after performing a Mann-Whitney U test to compare the samples of each class. This is a feature selection method that has been previously used in medical applications [117]. In addition, we took into account the efficiency of this feature selection approach, given the large feature space considered (Table 5.1). These models are defined as: the *selected features' model 1* with the features where $p<.05$, when comparing the suicidal ideation risk and control groups; and the *selected features' model 2*, where $p<.001$. The number of features obtained for each model is also given in Table 5.1.

It is relevant to mention that this has not been the only feature selection approach evaluated in this thesis for suicidality detection. In [136] we address a task of suicide risk assessment. The goal was to calculate the level of risk of users of committing suicide. We modeled the task as a multiclass classification problem, and applied a document classification approach, where X^2 [94] is used as a feature selection method.

5.2.3 Experimental setup

We considered 3 different aspects to analyze: 1) the utility of having defined the SPV, as we believed that this would allow our models to focus on suicidality by reducing the noise provided by tweets that make no reference to the subject of interest; 2) the individual and combined contribution of the different aspects we analyzed: textual, relational, behavioral, and image-based information; and 3) the change in the performance of models that use 2 different types of control groups, one constituted by users that make use of vocabulary related to suicide (focused control) and another group of generic users who might not make use of these terms at all.

All posts from the full profile of the user were considered for baselines 1 and 2, whereas most of the features for our proposed models and combinations were extracted exclusively

using the SPV, except for some elements extracted from the user's tweets metadata and features such as the STTR, which required the usage of the posts from the full profile. For each task, 70% of all the instances were retained for training, and the remaining 30% (around 25 users per class) were left for testing purposes as unseen cases. To keep balanced instances from each class, we used stratification for these sets. In addition to these test sets, we also evaluated our best models over a sample of 200 users labeled as doubtful cases. This is done to verify if, as the human annotator, the models are capable of identifying most of these cases as users that are likely to be at risk.

The PowerTransformer class from Python's Scikit-learn library was used to transform the feature values to a normal distribution-like representation using Yeo-Johnson's [115]. To choose the best classifier, a 10-fold cross-validation process was followed over the training set with all the algorithms to evaluate. Afterward, the ones with the best performance were selected to perform a second 5-fold cross-validation along with a grid search to find the most suitable parameters for the classifier chosen.

We considered the precision (Pr), recall (R), F1 score (F1), accuracy, and area under the receiver operating characteristics curve (AUC-ROC) score denoted as AUC, which was the measure on which we based the parameter optimization of the grid search. The values for Pr, R, F1, and AUC corresponded to the suicidal ideation risk class, as it is our main class of interest. We reported on accuracy to analyze the performance of both classes. The results obtained by certain classifiers such as the CNNs were averaged results of multiple runs because of the randomness they can add.

5.2.4 Results

Table 5.2 presents the evaluation measure results for each task on the test sets. We reported the results for the best models, as described in Table 5.1, along with the baselines. We describe our results in terms of the following elements analyzed:

Table 5.2. Predictive tasks' results in terms of precision (Pr), recall (R), F1-score (F1), accuracy (Ac) and area under the curve (AUC).

Model	Suicidal ideation versus focused control group						Suicidal ideation versus generic control group					
	Pr	**R**	**F1**	**Ac**	**AUC**	**Classifier**	**Pr**	**R**	**F1**	**Ac**	**AUC**	**Classifier**
BoW model—full profile (baseline 1)	0.78	0.81	0.79	0.78	0.81	MLP	0.79	0.85	0.81	0.80	0.91	MLP
Embeddings model—full profile (baseline 2)	0.76	0.81	0.79	0.77	0.82	CNN	0.78	0.87	0.82	0.80	0.84	CNN
BoW model—SPV (baseline 3)	0.81	0.85	0.83	0.82	0.85	LR	0.80	**0.92**	0.86	0.84	0.89	MLP
Embeddings model—SPV (baseline 4)	0.79	0.85	0.82	0.80	0.83	CNN	0.77	0.87	0.82	0.80	0.82	CNN
SNPSY model	0.85	0.85	0.85	0.84	0.86	SVM	0.85	0.88	0.87	0.86	0.94	LR
(BoW + SNPSY) model	0.82	**0.88**	0.85	0.84	0.89	RF	0.85	0.88	0.87	0.86	0.94	LR
(Images + BoW) model	0.79	**0.88**	0.84	0.82	0.86	MLP	0.82	0.88	0.85	0.84	0.90	LR
(Images + SNPSY) model	**0.88**	0.85	**0.86**	**0.86**	0.91	SVM	0.88	0.88	**0.88**	**0.88**	0.94	LR
(Images + BoW + SNPSY) model 1	0.85	0.85	0.85	0.83	0.87	LR	0.85	**0.92**	**0.88**	**0.88**	0.92	MLP
(Images + BoW + SNPSY) model 2	**0.88**	0.81	0.84	0.84	**0.92**	SVM	0.85	0.88	0.87	0.86	0.94	LR
Selected features' model 1 (p<.05)	0.85	0.85	0.85	0.84	0.90	MLP	**0.91**	0.77	0.83	0.84	0.94	SVM
Selected features' model 2 (p<.001)	0.83	0.77	0.80	0.80	**0.92**	SVM	**0.91**	0.81	0.86	0.86	**0.95**	SVM

a) Short profile version definition results

As can be seen in Table 5.2, the definition of the SPVC is successful as the first filter for both the predictive tasks. Indeed, the BoW models trained exclusively on the SPV (baselines 3 and 4) outperformed baselines 1 and 2 for most of the measures on both tasks.

For these representations, tweets unrelated to the topic seem to introduce noise, as they generate a bigger feature space. In contrast, setting a high decision threshold for the classifier implies reducing the vocabulary for the BoW model that shall be generated over the SPV. This might reduce the performance of the model with the test data.

Regarding the CNN embedding models trained exclusively on the SPV, we can see that the model of task 1 obtains slightly better results compared with the baseline 2 model, whereas the results do not differ much for task 2.

In general, we observed a better performance with the SPV for BoW models. Therefore, the combinations evaluated take into account these text-based representations (BoW).

It is important to recall that for the focused control cases, after applying the SPVC with a decision threshold over 0.5, 4 users were left without an SPV because none of their tweets obtained a predicted probability over the threshold.

Considering that with higher thresholds, more focused control and generic users could be lost for training our next classifier, *0.5* is the threshold we kept for our further experiments. However, these results also showed that using SPVC reduces the number of control users with an SPV as the threshold value rises.

Based on these prior findings, the definition of the SPV is useful for discarding users who do not present tweets similar to those of the users at risk.

Initially, we found that focused control users were more easily discarded than generic users. However, this could be explained by the fact that the control users discarded might correspond to informative accounts such as newspapers, which we assumed to make use of certain terms referring to suicide in a way that does not make use of terms that imply a personal reference or opinion; therefore, the first classifier might find it easier to discard. In any case, this is a supposition

as we did not have further access to the writings of users after the annotation.

b) Combining models results

Regarding the methods considered for combining the types of features extracted, we can observe that when these types are evaluated independently from each other, each has a good accuracy, with the SNPSY model obtaining the best results.

For the combinations reported in Table 5.1 for the suicidal ideation risk versus focused control groups, we can observe that the models that use the 3 types of features do not significantly improve the results obtained by the SNPSY model. However, for *(Images + BoW + SNPSY) model 2*, we can see a 7% and 11% increase in the AUC score compared with baseline 3 and baseline 1, respectively, for the suicidal ideation versus focused control cases. The AUC difference of their ROC curves using the Delong method was p=.04, which is statistically significant, considering a 95% CI.

For task 2, the *(Images + BoW + SNPSY)* combination obtained results that improved baseline 1 for the suicidal ideation versus generic control task. For the *(Images + BoW + SNPSY) model 1*, we noticed a 4% increase in accuracy compared with baseline 3, and it increased to 8% compared with baseline 1.

There was also an increase in the AUC value of up to 4% with the selected features model 2. We also noticed the same measured results between the SNPSY model, the *(BoW + SNPSY) model*, and the *(Images + BoW + SNPSY) model 1*, implying that we might not improve the performance of the SNPSY model by adding other feature types. In fact, after conducting a Delong test to compare the ROC curves of these models with the baseline 1 model, we could not find significant differences, implying that their performance was not significantly different from the baseline in terms of the AUC measure for this task. However, this also implied that the use of the SNPSY features alone allowed us to have a model with a reduced number of features that performs as well as the BoW model with thousands of features.

Regarding the role of the images, we can see that when they are individually combined either with the BoW features or the SNPSY features, either the F score or the AUC score increases minimally compared with baseline 3.

As part of the experiments for this approach, it is necessary to mention that as some image scores were missing for a few users (up to 4 for each task), the approach considered to address this issue was to replace the scores by the mean of all the users except for the model where only a single score for each feature type was considered; for this case, the instances with missing values were removed.

In reference to the models with a set of selected features, we can notice that these models also outperform baselines 1 and 2 in terms of F1, accuracy, and AUC.

The *selected features' model 1* for both tasks outperformed baselines 3 and 4 on F1 and AUC. It should be noted that these models consider a reduced number of features compared with the baseline models, and the *(Images + BoW) model*, as they attempted to reduce the overfitting that the usage of thousands of features might imply.

c) Comparative results of tasks

When comparing the results of both tasks, we saw that the results obtained by the models to distinguish users at risk from generic control users were not that different from those trained over focused control users. However, we noticed higher levels of certainty for the models trained to compare users at risk and generic control users. This can be observed when comparing the AUC scores, which are always higher for the models of task 2. In fact, for this task, we can see that a high AUC score is already obtained by the baseline models, and it does not improve significantly with other models. This differs from task 1, where the feature combination is relevant for improving the certainty of the models compared with the baseline.

Figure 5.1 shows the top 10 most correlated features with the class for each task considering the features of *(Images + BoW + SNPSY) model 2*.

The most correlated features were given by textual elements such as the BoW model scores and lexicons. It is interesting to see that a behavioral feature as the median time between tweets is relevant for task 2.

We can also notice that for both tasks, self-references are relevant and that the usage of explicit suicide terms and health-related terms is relevant for task 2, as generic control users are not characterized by the usage of terms related to suicide.

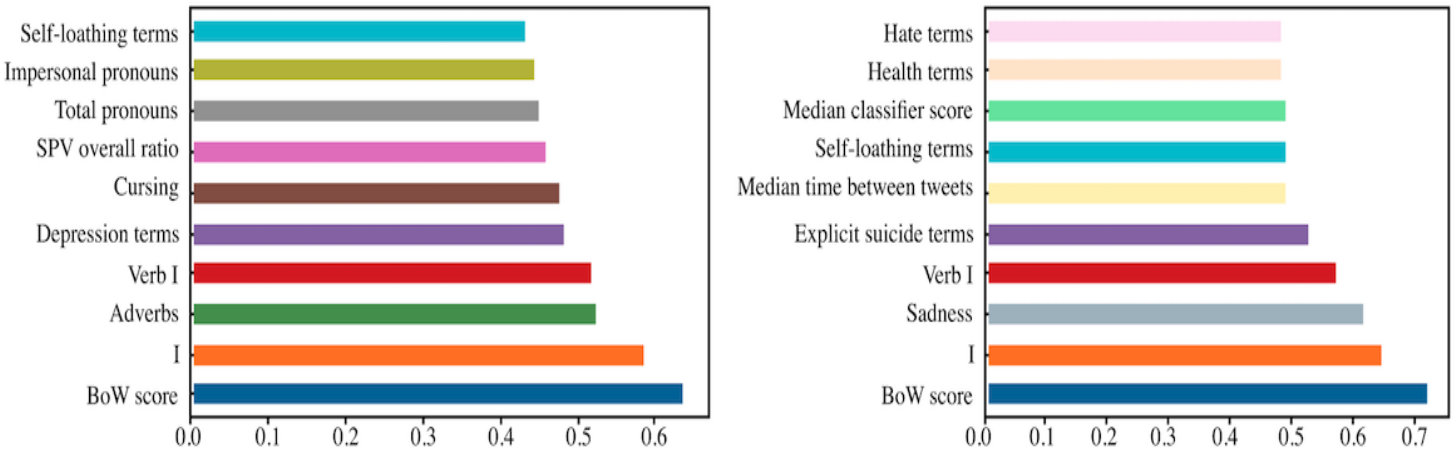

Figure 5.1. Features more correlated with the class to predict for both tasks: Suicidal ideation risk vs Focused control (left), and Suicidal ideation risk vs Generic control (right).

Referring to the features that were more predictive for the models generated, we used random forest's feature importance function, which is based on its measure of impurity. In this sense, we can see how much each feature decreases the impurity. The more a feature decreases the impurity the more important is the feature. In this case, because random forest uses multiple trees, the impurity decrease from each feature was averaged across all trees to determine the final importance of the variable. The most important features based on this approach, considering the features of *(Images + BoW + SNPSY) model 2*, are shown in Figure 5.2. For this case, we confirmed that for task 2, the usage of terms related to work and health is distinctive for both classes.

For both approaches, we can see that the image scores do not appear within the features more relevant for the tasks, implying that textual and behavioral features can be more relevant. Regardless of this, the scores given by certain feature combinations showed that the inclusion of the image scores improves minimally the results of these predictive tasks.

We also evaluated the selected features model 2, as one of our models with the best results for AUC for both tasks, over a sample of 200 users who were initially labeled as doubtful cases. We evaluated 2 models, one trained with the data of task 1 (*selected features' model 2*—task 1) and another trained with the data of task 2 (*selected features' model 2*—task 2). For the first model, we predicted 65% of the doubtful cases as positive (risk), whereas for the second model, 73% of the doubtful cases were found to be at risk. This indicates that our models detected signs of suicidal ideation in more than half of the doubtful users, which is in concordance with the criteria of the first annotator.

Finally, we evaluated the *selected features' model 2*—Task 1 over a test set of suicidal ideation and generic control users to evaluate the performance of this model over users who do not use a suicide-related vocabulary. We obtained the following results: Pr=0.91, R=0.77, F1=0.83, accuracy=0.84, and AUC=0.95. These results showed that the model obtains better results in comparison with its performance over focused control users. Similarly, we evaluated the *selected features' model 2*—task 2 over a test set of suicidal ideation and focused

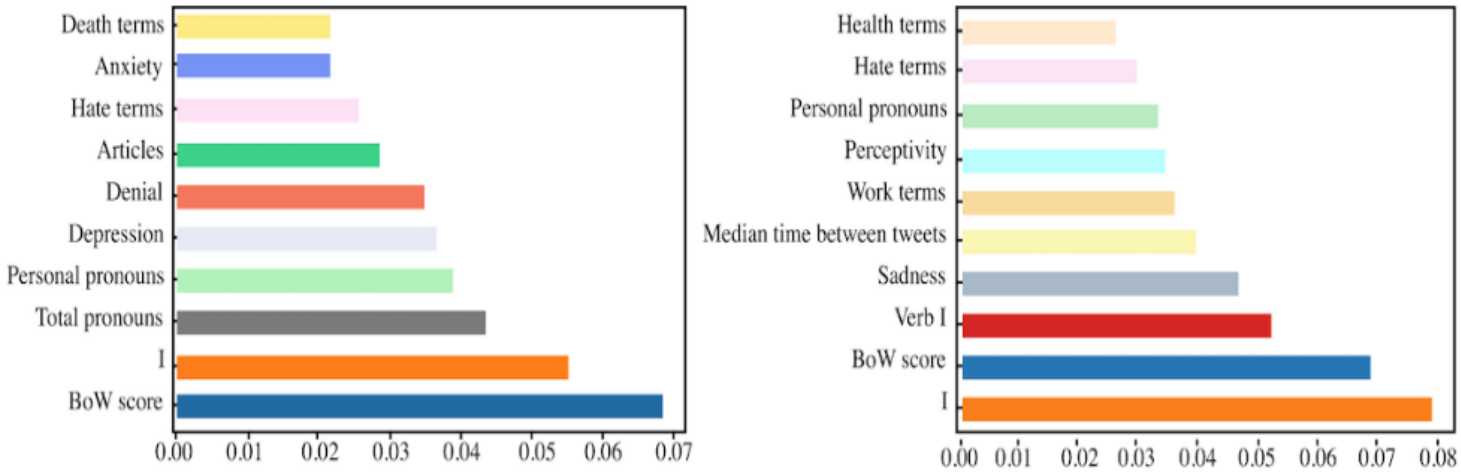

Figure 5.2. Most predictive features for both tasks: Suicidal ideation risk vs Focused control (left), and Suicidal ideation risk vs Generic control (right).

control users obtaining Pr=0.83, R=0.80, F1=0.82, accuracy=0.82, and AUC=0.91. The performance of this model was worse than that for generic control users. This is consistent with the fact that distinguishing these 2 cases is much harder.

5.3 Enhanced word embedding-based models

5.3.1 Introduction

We present a method for the generation of enhanced word embeddings for classification tasks on specialized domains. We present work dedicated to two binary tasks: 1) the detection of anorexia nervosa compared to control cases [128], and 2) the detection of mental disorders compared to control cases [127]. We also address a multiclass classification task dedicated to detect cases of eating disorders, suicidal ideation, alcoholism and depression.

For all the cases proposed we present an approach to generate word embeddings adapted to the domain of mental health assessment.

The first part of this section is dedicated to the detection of cases of anorexia nervosa using Dataset 1b – anorexia. We present a first approach for the enhancement of word embeddings adapting them to the classification task addressed.

The second part of this section presents an extension of the work presented in the first part, using dataset 4. This approach takes advantage of the method defined in section 4.2.3 dedicated to the detection of elements predictive for a given class, and generates word embeddings to address binary and multiclass predictive tasks.

The particularity of the tasks addressed is that texts are characterized by the usage of specific terms and expressions, like in the case of the eating disorders and alcoholism communities, where it is common to find terms as thinspiration, which refers to content that inspires a person to be thin; or AA, which is used to refer to Alcoholics Anonymous in the alcoholism community. Consequently, we consider that this specific vocabulary must be exploited to solve in a more efficient way the classification task. Notably we assume that the classical embedding models learnt on large generic

datasets are not suitable and that they must be adapted. Based on this hypothesis, our proposal takes advantage of the prior knowledge of terms that are predictive for each class, and generates representations suitable for the predictive task to address.

Researchers have created automated methods to detect mental disorders on social media by assuming that documents written by people presenting these disorders contain specific terms that describe signs and symptoms of a given condition [62]. However, before identifying these discriminant terms, it is necessary to find a suitable representation of the documents. Bag of Words (BoW) are among the most classical models considered. They allow to represent each text by a vector with components that are based on the number of times the terms of an index appear in the text. More recently, word embedding models have been introduced, and they have proved to be very efficient for solving text mining tasks. In these models, terms are represented by vectors that are generated under the principle that words appearing in similar contexts are related, and they should have close representations in the vector space. Thus, one can compute a similarity score between two words by calculating the cosine value of their corresponding word vector and, a high value indicates that they are semantically related.

Examples of methods developed to generate word embeddings models are: Word2vec [103], where a vector is generated for each word in the corpus considering it as an atomic entity; GloVe [116] that defines a weighted least squares model for training on global word-word co-occurrence counts; or fastText [21] that addresses the morphology of words in a way such that a term is represented as a bag of character n-grams. More recent methods have addressed the issue of generating context aware representations, where polysemic terms are taken into account. Instances of these types of representations are ELMo [118] and BERT [44].

Among the methods described, we consider Word2vec [103], and a distilled version of BERT: DistilBERT [139] to create a baseline model for one of our tasks. This method is selected as it has proved to generate models that are lighter and faster for fine-tuning purposes than BERT. The models

have also proved to be efficient enough in related tasks such as sentiment classification [139].

Embeddings that are generated through the prior approaches are often trained over large general corpora. However, when we consider their usage on domain specific classification tasks, in particular on the medical domain, it is common to have a reduced amount of labeled data to work with [63]. Moreover, embedding models learned exclusively in the domain corpus tend to not perform well on unseen cases with new vocabulary. Considering this issue, some methods have been developed to enhance the embeddings learned over small corpora. Those methods consist in incorporating external information [51], or adapting embeddings learned on large corpora to the task domain [39].

Within the enhancement methods, there is the work of [170] where approaches for combining different embedding sets to learn meta-embeddings are presented. Also, Faruqui *et al.* [51] propose a method that uses relational information from semantic lexicons for improving pre-built word vectors. Our approach surges as an alternative to handle small corpora and therefore some variations of these methods are considered as baselines to compare our model against other enhancement approaches.

We introduce a method based on *Dict2vec* [156], where in addition to the context defined by *Word2vec*, positive and negative sampling components are introduced. *Dict2vec* works by using the lexical dictionary definitions of words to enrich the semantics of the embeddings generated over small corpora. This approach is based on the fact that all the words in the definition of a term from a dictionary are semantically related to the word they define, and therefore, the positive sampling component moves closer the vectors of words co-occurring in their mutual dictionary definitions, and the controlled negative sampling prevents to move these vectors apart.

In our first approach, dedicated to a binary document classification task, the positive sampling component consists in moving close to each other the vector representations of terms that are predictive for the main target class by defining a pivot vector p towards which the vectors of predictive words are moved during the learning step. The negative sampling component, asides from preventing moving apart the vectors

of words that are predictive for the target class, also puts apart from *p* the vectors of the words that are the least predictive. In our second approach we add some modifications to this method through a modification of the objective function and the way to choose the set of words that are predictive for a given class, in a way such that enhanced embeddings can also be generated for multi-class classification tasks.

5.3.2 Anorexia detection

Our first method generates word embeddings enhanced for a classification task dedicated to the detection of users with AN over a small-sized corpus [128]. In this context, users are represented by documents that contain their writings concatenated, and that are labeled as anorexic (positive) or control (negative) cases (Dataset 1b - anorexia). These labels are known as the classes to predict for our task.

This method is based on Dict2vec's proposal [156]. We extend the Word2vec model with both a positive and a negative component, but our method differs from Dict2vec because both components are designed to learn vectors for a specific classification task. Within the word embeddings context, we assume that word-level n-grams' vectors, which are predictive for a class, should be placed close to each other given their relation with the class to be predicted. Therefore we first define sets of what we call predictive pairs for each class, and use them later for our learning approach.

The main contributions of this work are: 1) a method that modifies *Dict2vec* [156] in order to generate word embeddings enhanced for our classification tasks (binary and multiclass), this method has the power to be applied on similar tasks that can be formulated as document categorization problems; 2) different ways to improve the performance of the embeddings generated by our method corresponding to a set of embeddings variants; and 3) a set of experiments to evaluate the performance of our generated embeddings in comparison to pre-learned embeddings, and other domain adaptation methods.

a) Predictive pairs definition

Prior to learning our embeddings, we use X^2 [94] to identify the predictive n-grams. This is a method commonly used for

feature reduction, being capable of identifying the most predictive features, in this case terms, for a classification task.

Based on the X^2 scores distribution, we obtain the n terms with the highest scores (most predictive terms) for each of the classes to predict (positive and negative). Later, we identify the most predictive term for the positive class denoted as t_1 or pivot term. Depending on the class for which a term is predictive, two types of predictive pairs are defined, so that every time a predictive word is found, it will be put close or far from t_1. These predictive pair types are: 1) positive predictive pairs, where each predictive term for the positive class is paired with the term t_1 in order to get its vector representation closer to t_1; and 2) negative predictive pairs, where each term predictive for the negative class is also paired with t_1, but with the goal of putting it apart from t_1.

To define the predictive terms for a binary classification task, we consider: the predictive terms defined by the X^2 method, AN related vocabulary (domain-specific) and the k most similar words to t_1 obtained from pre-learned embeddings, according to the cosine similarity. Like this, information coming from external sources that are closely related with the task could be introduced to the training corpus. The terms that were not part of the corpus were appended to it, providing us an alternative to add new vocabulary of semantic significance to the task.

Regarding the negative predictive terms, no further elements are considered asides from the (X^2) predictive terms of the negative class as for our use case and similar tasks, control cases do not seem to share a vocabulary strictly related to a given topic. In other words, and as observed for the anorexia detection use case, control users are characterized by their discussions on topics unrelated to anorexia.

For the X^2 method, when having a binary task, the resulting predictive features are the same for both classes (positive and negative). Therefore, we have proceeded to get the top n most predictive terms based on the distribution of the X^2 scores for all the terms. Later, we decided to take a look at the number of documents containing the selected n terms based on their class (anorexia or control). Given a term *t*, we calculated the number of documents belonging to the positive class (anorexia) containing *t*, denoted as PCC; and we also calculated the number of documents belonging to the negative

class (control) containing *t*, named as NCC. Then, for *t* we calculate the respective ratio of both counts in relation to the total amount of documents belonging to each class: total amount of positive documents (TPD) and total amount of negative documents (TND), obtaining like this a positive class count ratio (PCCR) and a negative class count ratio (NCCR).

For a term to be part of the set of positive predictive terms its PCCR value has to be higher than the NCCR, and the opposite applies for the terms that belong to the set of negative predictive pairs. The positive and negative class count ratios are defined in Equations 5.1 and 5.2 as:

$$PCCR(t) = \frac{PCC(t)}{TPD} \quad (5.1) \qquad NCCR(t) = \frac{NCC(t)}{TND} \quad (5.2)$$

b) Learning embeddings

Given the positive and negative pairs, the aim of this method consists in determining a vector representation of the terms in such a way that the vectors of positive predictive terms are represented close to their corresponding pivot vector. These embedding representations are obtained by optimizing a global objective function. Adopting the notation in [156], the objective function for a target term ω_t (Equation 5.3) is given by the aggregation of Word2vec's (target term ω_t, context term ω_c) pair cost, a positive sampling cost (Equation 5.4) and a negative sampling cost (Equation 5.5). *Word2vec*'s cost is given by $\ell(v_t, v_c)$ where ℓ corresponds to the logistic loss function, and (v_t) and (v_c) are the vectors of ω_t and ω_c respectively.

$$J(\omega_t, \omega_c) = \ell(v_t, v_c) + J_{pos}(\omega_t) + J_{neg}(\omega_t) \quad (5.3)$$

The positive sampling component J_{pos} is calculated for each target term according to Equation 5.4:

$$J_{pos}(\omega_t) = \beta_P \sum_{\omega_i \epsilon P(\omega_t)} \frac{\ell(v_t \,.\, v_i)}{|P(\omega_t)|} \quad (5.4)$$

$P(\omega_t)$ represents the set of n-grams that form a positive predictive pair with the n-gram ω_t. The vectors v_t and v_i represent ω_t and ω_i respectively. Like in Dict2vec, a weight β_P represents the importance of the positive sampling component during the learning phase. The cost given by the predictive pairs is normalized by the size of the predictive pairs set, $|P(\omega_t)|$, considering that all the terms from the predictive pairs set of ω_t are taken into account for the calculations, and therefore when t1 is found, the impact of trying to move it closer to a large quantity of terms is reduced, and it remains as a pivot element to which other predictive terms get close to.

For the negative sampling cost J_{neg} defined in Equation 5.5, according to the first component, the vectors of the terms forming a positive predictive pair with ω_t are not moved away from ω_t thanks to the modification of the negative random sampling cost of *Word2vec*, where a set $F(\omega_t)$ of k random terms from the vocabulary are moved away from the vector of ω_t considering that those random terms are not likely to be semantically related. We not only make sure that the vectors of the terms forming a positive predictive pair with ω_t are not put apart from it, but in the second component of Equation 5.5 define a cost given by the negative predictive pairs. In this case, as explained before, the main goal is to put apart terms that are not predictive for the main class from t_1, so this cost is added to the negative random sampling cost. In this case, $N(\omega_t)$ represents the set of all the words that form a negative predictive pair with the word ω_t. β_N represents the weight that defines the importance of the negative component.

$$J_{neg}(\omega_t) = \sum_{\substack{\omega_i \epsilon F(\omega_t) \\ \omega_i \notin P(\omega_t)}} \ell(-v_t \,.\, v_i) + \beta_N \sum_{\omega_j \epsilon N(\omega_t)} \frac{\ell\left(-v_t \,.\, v_j\right)}{|N(\omega_t)|} \qquad (5.5)$$

Finally, the sum of the cost of every (target, context) pair is what defines the global objective function (Equation 5.6) where n is the size of the window and C is the corpus size.

$$J = \sum_{t=1}^{C} \sum_{c=-n}^{n} J(\omega_t, \omega_{t+c}) \qquad (5.6)$$

c) Embeddings variations

Given a pre-learned embedding which associates for a word ω a pre-learned representation v_{pl}, and an enhanced embedding v obtained through our approach for ω with the same length m as v_{pl}, we generate variations of our embeddings based on existing enhancement methods. First, we denote the embeddings generated exclusively by our approach (predictive pairs) as *Variation 0*, v is an instance of the representation of ω for this variation.

For the next variations, we address ways to combine the vectors of pre-learned embeddings (*i.e.*, v_{pl}) with the ones of our enhanced embeddings (*i.e.*, v). For *Variation 1* we concatenate both representations ($v_{pl} + v$), obtaining a $2m$ dimensions vector [170]. *Variation 2* involves concatenating both representations and applying truncated SVD as a dimensionality reduction method to obtain a new representation given by $SVD(v_{pl} + v)$. *Variation 3* uses the values of the pre-learned vector v_{pl} as starting weights to generate a representation using our learning approach. This variation is inspired by a popular transfer learning method that was successfully applied on similar tasks [39]. For these variations (1-3) we take into account the intersection between the vocabularies of both embeddings types (pre-learned and *Variation 0*). Finally, *Variation 4* implies applying Faruqui's retrofitting method [51] over the embeddings of *Variation 0*.

d) Evaluation framework

We used Dataset 1b – anorexia (see Table 3.3) that consists of posts of users labeled as anorexic and control cases. Given the incidence of anorexia nervosa, for both sets there is a reduced yet significant amount of AN cases compared to the control cases. Here we address our embeddings generation approach, and the evaluation methods defined along with their respective results.

- Embeddings generation

The training corpus used to generate the embeddings, named anorexia corpus, consisted on the concatenation of all the writings from all the training users. A set of stop-words were removed. This resulted in a training corpus with a size of 1,267,208 tokens and a vocabulary size of 87,197 tokens. In order to consider the bigrams defined by our predictive pairs,

the words belonging to a bigram were paired and formatted as if they were a single term.

For the predictive pairs' generation with X^2, each user is an instance represented by a document composed by all the user's posts concatenated. X^2 is applied over the train set considering the users classes (anorexic or control) as the possible categories for the documents. The process described to define the predictive pairs is followed in order to obtain a list of 854 positive (anorexia) and 15 negative (control) predictive terms. Some of these terms can be seen on Table 5.3, which displays the top 15 most predictive terms for both classes. The term anorexia was the one with the highest X^2 score, denoted as t1 in the predictive pairs' definition.

Table 5.3. List of some of the most predictive terms for each class.

Positive Terms (Anorexia class)			**Negative terms (Control class)**		
• anorexia	• diagnosed	• binges	• war	• sky	• song
• anorexic	• macros	• calories don't	• bro	• plot	• master
• meal plan	• cal	• relapsed	• Trump	• game	• Russian
• underweight	• weight gain	• restriction	• players	• Earth	• video
• eating disorder(s)	• anorexia nervosa	• caffeine	• gold	• America	• trailer

The anorexia domain related terms from [11] were added as the topic related vocabulary, and the top 20 words with the highest similarity to anorexia coming from a set of pre-learned embeddings from GloVe [116] were also paired to it to define the predictive pairs sets. The GloVe's pre-learned vectors considered are the 100 dimensions representations learned over 2B tweets with 27B tokens, and with 1.2M terms.

The term anorexia was paired to 901 unique terms and, likewise, each of these terms was paired to anorexia. The same approach was followed for the negative predictive terms (15), which were also paired with the pivot term anorexia. An instance of a positive predictive pair is (anorexia, underweight), whereas an instance of a negative predictive pair is (anorexia, game). For learning the embeddings through our approach, and as it extends Word2vec, we used as parameters a window size of 5, the number of random negative pairs chosen for

negative sampling was 5, and we trained with one thread/worker and 5 epochs.

- Evaluation based on the average cosine similarity

This evaluation is done over the embeddings generated through *Variation 0* over the anorexia corpus. It averages the cosine similarities (sim) between t_1 and all the terms that were defined either as its *p* positive predictive pairs, obtaining a positive score denoted as PS on Equation 5.7a; or as its n negative predictive pairs, with a negative score denoted as NS on Equation 5.7b. On these equations v_a represents the vector of the term *anorexia*; v_{PPT_i} represents the vector of the positive predictive term (PPT) *i* belonging to the set of positive predictive pairs of *anorexia* of size *p*; and v_{NPT_i} represents the vector of the negative predictive term (NPT) *i* belonging to the set of negative predictive pairs of *anorexia* of size *n*:

$$PS(a) = \sum_{i=1}^{p} \frac{sim(v_a, v_{PPT_i})}{p} \quad (5.7a) \qquad NS(a) = \sum_{i=1}^{n} \frac{sim(v_a, v_{NPT_i})}{n} \quad (5.7b)$$

We designed our experiments using PS and NS in order to analyze three main aspects: 1) we verify that through the application of our method, the predictive terms for the positive class are closer to the pivot term representation, and that the predictive terms for the negative class were moved away from it; 2) we evaluate the impact of using different values of the parameters β_P and β_N to obtain the best representations where PS has the highest possible value, keeping NS as low as possible; and 3) we compare our generation method with *Word2vec* as baseline since this is the case for which our predictive pairs would not be considered $\beta_P = 0$ and $\beta_N = 0$).

We expect our embeddings to obtain higher values for PS and lower values for NS in comparison to the baseline.

Table 5.4 shows first the values for PS and NS obtained by what we consider our baseline, *Word2vec* $\beta_P = 0$ and $\beta_N = 0$), and then the values obtained by embeddings models generated using our approach *Variation 0*, with different yet equivalent values given to the parameters β_P and β_N, as they proved to provide the best results for PS and PN. After applying our approach, the value of PS becomes greater than NS for

Table 5.4. Positive cores (PS) and Negatives cores (NS) for Variation 0. Different values for β_P and β_N are tested.

values for β_P and β_N	Score positive (SP)	Score negative (SN)
β_P = 0, β_N = 0 (word2vec)	0.89	0.90
β_P = 0.25, β_N = 0.25	0.79	0.74
β_P = 0.5, β_N = 0.5	0.79	0.52
β_P = 1, β_N = 1	0.80	0.59
β_P = 10, β_N = 10	0.85	**0.47**
β_P = 50, β_N = 50	**0.95**	0.60
β_P = 100, β_N = 100	0.93	0.64

most of our generated models, meaning that we were able to obtain a representation where the positive predictive terms are closer to the pivot term *anorexia*, and the negative predictive terms are more apart from it.

Then, we can also observe that the averages change significantly depending on the values of the parameters β_P and β_N, and for this case the best results according to PS are obtained when β_P = 50 and β_N = 50. Finally, when we compare our scores with *Word2vec*, we can observe that after applying our method, we can obtain representations where the values of PS and NS are respectively higher and lower than the ones obtained by the baseline model.

- Evaluation based on Visualization

We focus on the comparison of embeddings generated using *Word2vec* (baseline), *Variation 0* of our enhanced embeddings, and *Variation 4*. In order to plot over the space the vectors of the embeddings generated (see Figure 5.3), we performed dimensionality reduction, from the original 200 dimensions to 2, through Principal Component Analysis (PCA) over the vectors of the terms in Table 5.3 for the embeddings generated with these three representations. We focused over the embeddings representing the positive and negative predictive terms. For the resulting embeddings of our method *Variation 0* we selected β_P = 50 and β_N = 50 as parameter values. The positive predictive terms representations are closer after applying our method *Variation 0*, and the negative predictive terms are displayed farther, in comparison to the baseline.

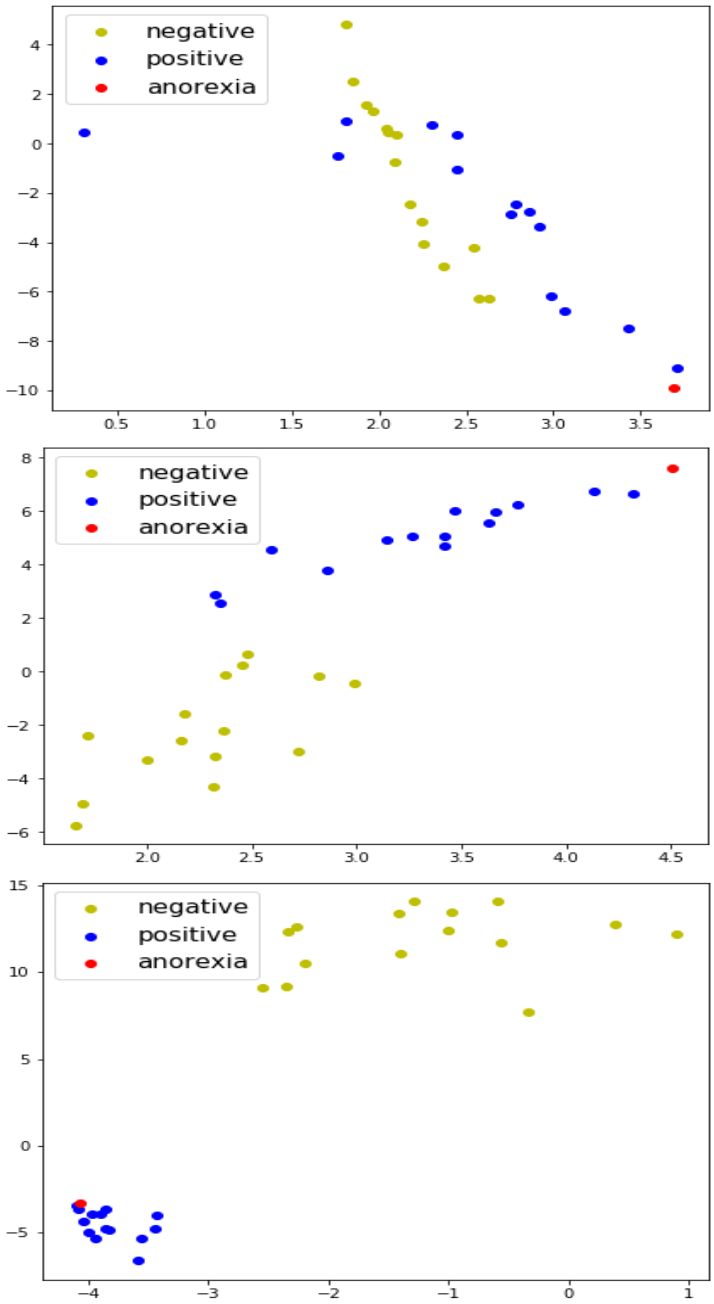

Figure 5.3. Predictive terms sample represented in two dimensions after PCA was applied on their embeddings as dimensionality reduction method. From top to bottom each plot shows the vector representation of the predictive terms according to the embeddings obtained through 1) Word2vec (baseline), 2) Variation 0, and 3) Variation 4.

The last plot (bottom) displays the terms for the embeddings generated through *Variation 4*. For this case, given the input format for the retrofitting method, *anorexia* was linked with all the remaining predictive terms of the anorexia class (901), and likewise, each of these predictive terms was linked to the term *anorexia*.

Notice that the retrofitting approach converges to changes in Euclidean distance of adjacent vertices, whereas the closeness between terms for our approach is given by the cosine distance.

- Evaluation based on a predictive task

In order to test our generated embeddings for the classification task dedicated to AN screening, we conduct a series of experiments to compare our method with related approaches.

We define 5 baselines for our task: the first one is a BoW model based on word level unigrams and bigrams (*Baseline 1)*, this model is kept mainly as a reference since our main focus is to evaluate our approach compared to other word embedding based models. We create a second model using *GloVe's* pre-learned embeddings (*Baseline 2)*, and a third model that uses word embeddings learned on the training set with the *Word2vec* approach (*Baseline 3)*. We evaluate a fourth approach (*Baseline 4)* given by the enhancement of the (*Baseline 3)* embeddings, with Faruqui's *et al.* [51] retrofitting method. *Baseline 5* uses the same retrofitting method over GloVe's pre-learned embeddings, as we expected that a domain adaptation of the embeddings learned on an external source could be achieved this way.

To create our predictive models, again, each user is an instance represented by their writings. For *Baseline 1* we did TF-IDF vectorization of the users' documents, by using the TfIdfVectorizer provided by the Scikit-learn Python library, with a stop-words list and the removal of the n-grams that appeared in less than 5 documents. The representation of each user through embeddings was given by the aggregation of the vector representations of the words in the concatenated texts of the users, normalized by the size (words count) of the document. Then, an L_2 normalization was applied to all the instances.

Given the reduced amount of anorexia cases on the training set, we used SMOTE [28] as an over-sampling method to deal with the unbalanced classes. The Scikit learn's Python library implementations for Logistic regression (LR), Random Forest (RF), Multilayer Perceptron (MLP), and Support Vector Machines (SVM) were tested as classifiers over the training set with a 5-fold cross validation approach. A grid search over each method to find the best parameters for the models was done.

The results of the baselines are compared to models with our variations. For *Variation 4* and *baselines 4* and *5* we use the 901 predictive terms described for the visualization evaluation.

To define the parameters of *Variation 3*, we test different configurations, and chose the ones with the best results according to PS.

Precision (P), Recall (R), F1-Score (F1) and Accuracy (A) are used as evaluation measures. The scores for P, R and F1 reported over the test set on Table 5.5 correspond to the anorexia (positive) class, as this is the most relevant one, whereas A corresponds to the accuracy computed on both classes.

Seeing that there are 6 times more control cases than AN and that false negative (FN) cases are a bigger concern compared to false positives, we prioritize R and F1 over P and A. This is done because as with most medical screening tasks, classifying a user at risk as a control case (FN) is worse than the opposite (FP), in particular on a classifier that is intended to be a first filter to detect users at risk. Table 5.5 shows the results for the best classifiers. The best scores are highlighted for each measure.

Comparing the baselines, we can notice that the embeddings based approaches provide an improvement on R compared to the BoW model, however this is given with a significant loss on P.

Regarding the embeddings based models, the variations outperform the results obtained by the baselines. The model with the embeddings generated with our method (*Variation 0*) provides significantly better results compared to the *Word2vec* model (*Baseline 3)*, and even the model with pre-learned embeddings (*Baseline 2*), with a wider vocabulary.

Table 5.5. Baselines and enhanced embeddings evaluated in terms of precision (P), recall (R), F1-score (F1) and Accuracy (A).

Model	Description	P	R	F1	A	Classifier
Baseline 1	BoW Model	**90.00%**	65.85%	76.06%	**94.69%**	MLP
Baseline 2	GloVe's pre-learned embeddings	69.57%	78.05%	73.56%	92.81%	MLP
Baseline 3	Word2vec embeddings	70.73%	70.73%	70.73%	92.50%	SVM
Baseline 4	Word2vec retrofitted embeddings	71.79%	68.29%	70.00%	92.50%	SVM
Baseline 5	GloVe's pre-learned embeddings retrofitted	67.35%	**80.49%**	73.33%	92.50%	MLP
Variation 0	Predictive pairs embeddings (β_P = 50, β_N = 50)	77.50%	75.61%	76.54%	94.03%	MLP
Variation 1	Predictive pairs embeddings (β_P = 50, β_N = 50) + GloVe embeddings	69.57%	78.05%	73.56%	92.81%	MLP
Variation 2	Predictive pairs embeddings (β_P = 50, β_N = 50) + GloVe embeddings	75.00%	**80.49%**	**77.65%**	94.06%	MLP
Variation 3	Predictive pairs embeddings + GloVe embeddings starting weights (β_P = 0.25, β_N = 50)	72.73%	78.05%	75.29%	93.44%	MLP
Variation 4	predictive pairs (β_P = 50, β_N = 50) retrofitted embeddings	82.86%	70.73%	76.32%	**94.37%**	SVM

The combination of pre-learned embeddings and embeddings learned on the training set, provide the best results in terms of F1 and R. They also provide a good accuracy considering that most of the test cases are controls. Using the weights of pre-learned embeddings *Variation 3* to start the learning process over the corpus significantly improves the R score in comparison to *Word2vec*'s generated embeddings (*Baseline 3)*.

Finally, the worst results for the variations are given by *Variation 1* that obtains equivalent results to *Baseline 2*. The best model in terms of F1 corresponds to *Variation 2*. Also, better results are obtained for P when the embeddings are enhanced by the retrofitting approach *Variation 4*.

5.3.3 Detection of multiple mental disorders

We extend the work described in the prior section (5.3.2) and improve our embeddings' generation approach for addressing binary predictive task. We also introduce a way to adapt the method for it to be suitable to multiclass predictive tasks [127]. Using dataset 4, we formalize the problem in two ways: 1) as a multi-class classification task (Task 1) dedicated to the detection of posts related to depression (DEP), eating disorders (ED), suicidal ideation (SUI) and alcoholism (ALC) (Dataset 4a – multiple); and 2) as a binary classification task (Task 2) dedicated to the classification of posts (texts) of users with self-references related to substance abuse and mental health issues (MEN), and control posts (CON) which do not make reference to any of the prior conditions (Dataset 4b – mental).

Our proposal takes advantage of the prior knowledge of terms that are predictive for each class and generates representations suited for the predictive task to address. It extends the work described in the prior section (5.3.2), where enhanced word embeddings are generated for a binary classification task dedicated to anorexia nervosa screening.

The main new contributions of this work are: 1) a method that improves and adapts the model presented in Section 5.3.2 to address a multi-class classification task; 2) the creation of predictive models based on deep learning approaches to compare our enhanced representations against other word

embeddings' learning methods and domain adaptation approaches; 3) a type of feature designed for predictive models (named PSim) that leverages the properties of the embeddings generated with our method.

a) Embeddings generation

Compared to our prior approach (Section 5.3.2) this time we also address multiclass classification tasks. We represent the vectors of terms (word level n-grams) that are predictive for a given class close to each other, and far from those terms that are predictive for the remaining classes. For that, we also define positive and negative predictive pairs. These are based on the definition of a list of terms (word level {1-2}-grams) which are themselves predictive for each class. These pairs are later used as inputs of the embedding learning model.

- Predictive pairs generation – multiclass predictive task

To provide an appropriate input to our learning model, we define a set of positive predictive pairs and a set of negative predictive pairs for each type of task (multi-class or binary).

In order to address a multiclass predictive task, for each class c_n, a list of *positive predictive terms* denoted as c_n*_predictive* is built. To do so, we use the process dedicated to detect relevant n-grams previously described in Section 4.2.3 – Algorithm 1. Once we have a list of predictive terms for each class, we proceed to generate a list of negative predictive terms for each class. The list of *negative predictive terms* of c_n, which is denoted as c_n*_negative_predictive,* contains all the terms that are part of the list of positive predictive terms of every other class.

Once the lists of positive and negative predictive terms ({1-2}-grams) of each class are defined, we proceed to generate the inputs required for our embeddings learning approach, which consist of two lists of predictive pairs: the positive predictive pairs' list and the negative predictive pairs' list.

To generate the list of positive predictive pairs, we select one pivot term for each class in the list of classes C. The pivot term for a class c_n is given by the term with the highest X^2 score within the terms in c_n_predictive. This will be considered as a pivot term and the vectors of the predictive terms of c_n will be moved towards the vector of this term. Considering our use case as an instance (task 1), the pivot terms for the suicide, depression, eating disorders, and alcoholism classes were

respectively: *kill, depression, eating,* and *alcoholism*. A positive predictive pair is then composed by a pivot term and a term that is part of the list of positive predictive terms of the class for which the pivot term belongs to. Considering our use case (task 1), positive predictive pairs instances are (*eating, anorexia*) and (*alcoholism, beer*). This approach consists in pairing with their pivot term all the terms of the list of positive predictive terms of each class to compose the corresponding positive predictive pairs list.

For generating the list of negative predictive pairs, each pair is given by a pivot term, and a term that belongs to the list of negative predictive terms of the class for which the pivot term belongs to.

In our use case, examples of negative predictive pairs' instances are *(eating, beer)*, and *(alcoholism, anorexia)*, as this pairing approach consists in pairing with their pivot term all the terms of the list of negative predictive terms for each class in *C*. Each of these pairs are added to the negative predictive pairs list.

- Predictive pairs generation – binary predictive task

For the case where there are only two classes (c_1 and c_2), we follow a similar approach as for the multi-class task, but we consider that for binary tasks, the X^2 resulting predictive terms are the same for both classes. In this sense, we differ from our prior approach (Section 5.3.2), which assigns the class for which a term is relevant based on the ratio of documents that contain the term and on the class they belong to. Instead, the predictive terms are obtained based on the approach described in the Algorithm 2 of Section 4.2.3, which was dedicated to detect predictive n-grams when 2 classes are compared.

Later, considering that we only address two classes, we define a single pivot term, which is given by the term that obtains the highest X^2 score, and that is part of the list of predictive terms of one of the classes to predict.

With our use case (*task 2*) as an instance, and considering its nature, where control cases are characterized by terms that are not related to mental disorders but that can be related to many other types of topics, we choose our pivot term (the word *feel*) using the main class to predict, which is the *MEN* class.

For this case, a positive predictive pair is composed by the pivot term, and a term that is part of the list of predictive terms of the pivot term's class. In our use case, instances of positive predictive pairs for *task 2* are (*feel, abused*), (*feel, antidepressants*) and (*feel, attempted suicide*).

Finally, each negative predictive pair is composed by the same single pivot term (*feel*), and a term that is part of the predictive terms list of the remaining class. For our use case, instances of negative predictive pairs are (*feel, account*), (*feel, mechanical*), or (*feel, agent*).

- Learning approach

Given the positive and negative pairs, the aim of this method consists in determining a vector representation of the terms in such a way that the vectors of positive predictive terms are represented close to their corresponding pivot vector and far from the pivot vectors of the remaining classes.

These embedding representations are obtained by optimizing a global objective function. The function is the same as Equation 5.3 with variations of the positive (Equation 5.4) and negative sampling costs (Equation 5.5) presented in Section 5.3.2.

As it is our goal to keep the vector of the pivot term as a fixed element towards which other predictive terms get close to, whenever a pivot term happens to be the target term, the positive and negative sampling values are null. In this sense the positive sampling cost is zero if ω_t is a pivot term, otherwise its value is calculated according to Equation 5.8. This represents a modification of the cost considered in our prior approach (Equation 5.4) where the same issue of keeping the pivot term as fixed as possible was addressed by normalizing the cost with the size of the predictive pairs set of the term (|P(ω_t)|).

$$J_{pos}(\omega_t) = \beta_P \sum_{\omega_i \epsilon P(\omega_t)} \ell(v_t \,.\, v_i) \tag{5.8}$$

The negative sampling cost is given by Equation 5.9. We retake the elements and notation of Equation 5.5 and as for the positive sampling cost, to not affect the position of the vectors of pivot terms, whenever ω_t is a pivot term, the cost of the second component in (Equation 5.9) is zero.

$$J_{neg}(\omega_t) = \sum_{\substack{\omega_i \epsilon F(\omega_t) \\ \omega_i \notin P(\omega_t)}} \ell(-v_t \,.\, v_i) \; + \beta_N \sum_{\omega_j \epsilon N(\omega_t)} \ell\left(-v_t \,.\, v_j\right) \qquad (5.9)$$

Finally the global objective function remains as how it was described in Equation 5.6.

- Enhanced embeddings variations

For both use cases, we define 4 variations of our embeddings with the aim of improving the representations obtained. For this purpose we consider related approaches [39,51,170] with which our method is compatible.

We label our proposed model as *Embedding model 0*. Thus, the first variation (*Embedding model 1*) consists in learning embeddings with our model after using GloVe's pre-learned embeddings to define the starting weights of the vectors of terms. The second variation (*Embedding model 2*) consists in applying Faruqui's retrofitting method [51] over the representations of the *Embedding model 0*.

For the third variation, given a pre-learned embedding that associates for a term ω a pre-learned vector v_{pr}, and a vector v learned through our approach for the same term ω with the same length n as v_{pr}, an embedding of the *Embedding model 3* is defined by the concatenation of both representations $(v_{pr} + v)$ and the application of truncated SVD as a dimensionality reduction method so that the new vector of ω is given by $SVD(v_{pr} + v)$ [170], this variation is considered because it obtained the best results for the binary classification task addressed in Section 5.3.2.

Finally, the *Embedding model 4* corresponds to the retrofitting approach applied over the *Embedding model 1*.

- Features for predictive models based on enhanced embeddings

We propose a feature generation method that leverages the properties of the embeddings generated through our method. The obtained features can be used for machine learning models.

Our proposal takes into account that in our embedding model, the predictive terms of a class *c* are represented close to its pivot term in the vector space. Thus, if we define a vector

representation of a writing (document) that corresponds to *c*, and consider that predictive terms of *c* are likely to be found in the writing, we can assume that their presence is likely to influence the placement of the vector that represents the whole writing. In this sense, the vector that represents the document (a writing/post) should be closer to the vector of the pivot term of *c* in comparison to the vectors of documents that do not contain the predictive terms.

Based on our prior statement, we define the pivot similarity (*PSim*), which is calculated for each document and for each class to predict. Considering c as a class from the set of classes to predict *C*, t a term belonging to the set *T* of *n* terms composing the document *D*, v_t being the vector representation of *t* and, vp_c representing the vector of the pivot term of *c*, the value of *PSim* for *D* and *c* is defined by the cosine similarity between vp_c and the average of the vectors associated to the terms belonging to *D*. It is given by Equation 5.10.

$$PSim(D, c) = \cos\_sim\left(\frac{\sum_{t \in T} v_t}{n}, vp_c\right) \tag{5.10}$$

In a document classification task, for each document there will be as many features as classes to predict, except for a binary task, where as there is a single pivot term there is only one feature to define. Each feature corresponds to the $PSim$ value between the document and the pivot term of a class.

b) Embeddings experimental and evaluation framework

Here we explain the embedding generation process in our use cases dedicated to the detection of writings related to mental disorders. We also describe the methods adopted to evaluate the embeddings generated as well as their variations, and the results of these evaluations.

- Embeddings generation process

In order to generate the embeddings and to evaluate their performance, dataset 4a-multiple and dataset 4b-mental were split into training (70%) and test sets (30%). The distribution of the instances on each split was proportional for each class.

Table 5.6 gives the details of the datasets for both tasks. Two corpora were defined, a corpus corresponding to dataset 4a-multiple and a corpus that corresponds to dataset 4b-mental.

Table 5.6. Train and test sets description.

Task	Class	Train set Number of posts	Test set Number of posts
Task 1 (dataset 4a-multiple)	SUI	5,306	1,769
	DEP	2,261	754
	ALC	188	62
	ED	588	196
Task 2 (dataset 4b-mental)	MEN	8,343	2,781
	CON	15,042	5,015

The process defined to generate the predictive pairs was applied over the training set, where each post was represented by a document and its label, which corresponds to the document class (SUI, DEP, ED, ALC, MEN, or CON). Then, we generated the predictive pairs following the approach proposed.

Table 5.7 shows the list of the top 15 most predictive terms for each class. For the alcoholism class, we observed that despite having a reduced number of posts (see Table 5.6) it is a class that can be characterized by a large number of terms, whereas the suicide class, despite having the largest number of writings, does not have a large amount of unique distinguishable terms. Also, for this same task, the list of negative predictive terms for each class was given by the list of terms that were predictive for all the other classes.

For task 2, the number of positive predictive pairs obtained was 351, and the number of negative predictive pairs was 202. Table 5.8 shows the top 15 most predictive terms for each class.

From the pre-learned embeddings of GloVe [116], we also consider the top 20 terms with the highest similarity to each pivot term (eating, kill, depression and alcoholism for task 1, and feel for task 2) and terms highly related to the conditions such as *anorexia, suicide, bulimia, die, anxiety, eating disorder, alcohol* and *alcoholic*.

We add the terms to the list of predictive terms of the respective class only if they are relevant for the class according to the X^2 score, or if they are not already part of the vocabulary and are semantically related to the pivot term considering the context of the task.

Table 5.7. Pivots and list of the top 15 most predictive terms for each class (task 1).

Class	SUI	DEP	ED	ALC
Pivot terms	Kill	Depression	Eating	Alcoholism
Terms' number	11	6	45	56
Terms	Suicide	Anxiety	Eating disorder	Alcohol
	Die	Depressed	Bulimia	Alcoholic
	Want die	Depression anxiety	Purging	Drinking
	Killing	Energy	Ed	Drink
	Live	Mental health	Purge	Sober
	Dead	Sad	Weight	AA
	Anymore	-	Recovery	Beer
	Just want	-	Food	Sobriety
	Cares	-	Anorexia	Drank
	Care	-	Eat	Liquor
	Kill_myself	-	Binge	Drinks
	-	-	Calories	Drunk
	-	-	Bulimic	Stop drinking
	-	-	Binging	Beers
	-	-	Restricting	Drinking problem

Table 5.8. Pivot and list of the top 15 most predictive terms for each class (task 2).

MEN (pivot: *feel*)	**CON**
life	company
kill	customer
depression	calls
die	theory
friends	engineering
suicide	information
depressed	service
suicidal	book
feeling	legal
mental	number
anxiety	center
hate	question
pain	phone
shit	engineer
scared	account

This last aspect is considered because of terms such as *die* for instance, as most of the vectors of terms with the highest cosine similarity correspond to words in German. In the case where a term was not part of the corpus vocabulary, prior to learning, it was added to the corpus at the end of the last document belonging to the class. The GloVe's pre-learned vectors were the 100 dimensions embeddings learned over 2B tweets with 27B tokens, and with 1.2M vocabulary terms. These embeddings are also the ones used for the baselines and some of our embedding variations.

After having the predictive pairs defined, in order to generate the embeddings, for their corresponding task, each corpus considered for training purposes consisted of the concatenation of all the texts from all the training posts. Stop words were removed. This resulted in a training corpus with a size of 800,319 tokens and a vocabulary size of 23,450 unique terms for dataset 4a-multiple, and a corpus with a size of 2,230,423 tokens, with a vocabulary of 55,620 unique terms for dataset 4b-mental. For both datasets, to consider the predictive bigrams on the learning process, the words forming a predictive bigram were represented as a single term in the corpus. To learn our embeddings, we used as hyper parameters a window size of 5 with 5 random negative pairs chosen for negative sampling. We trained with one thread per worker and 5 epochs. Different values for β_P and β_N were tested.

- Evaluation approaches – average cosine similarity

We adapt the evaluation approaches of Section 5.3.2 that consist in a cosine similarity based evaluation, an evaluation based on visualization, and a predictive task evaluation.

The first evaluation approach is the *average cosine similarity evaluation*. It is applied over the *Embedding model 0.*

For the case of task 1, for each class *c* we average the cosine similarities between the vector of the pivot term of *c* and each of the vectors of the remaining positive predictive terms of *c* to obtain a positive score *P* for the class. We also calculate a negative score *N* for each class, which is given by the average of the cosine similarities between the vector of the pivot term of *c* and each of the vectors of the remaining negative predictive terms of *c*. To address task 2, *P* and *N* are calculated considering the pivot term (*feel*) of the main class to predict, which is the MEN class in this case.

For this evaluation approach we also choose as baselines: the embedding model that we introduced in section 5.3.2 (*Baseline 1*) where the positive and negative components of the objective function were normalized considering the size of the list of predictive terms for the target term; and the embedding model Word2vec exclusively (*Baseline 2*), which corresponds to the case where $\beta_P = 0$ and $\beta_N = 0$.

For the proposed models and, in comparison with the baselines, we expect to obtain better representations with our enhanced embeddings such a way that *P* keeps a high value while *N* is kept as low as possible. We also study the impact of the parameters by assigning different values to β_P and β_N.

For both tasks, the best results were obtained with equal values for β_P and β_N. For Task 1, they are described in Table 5.9. Remember that for *P* the higher the score the better, while for *N* the lower the better. Even though the values for *P* for most of the models are lower compared to *Baseline 2*, a good balance is obtained considering how the value for N decreases for all the classes. This means that the method has managed to obtain representations where the vectors of predictive terms for a class are represented far enough from the vectors of terms that are predictive for other classes, while keeping a high cosine similarity with the vectors of the terms that are predictive for their own class.

Notice that for *Baseline 1*, which corresponds to our prior method (Section 5.3.2), we present the configuration for $\beta_P = 10$ and $\beta_N = 10$ as it obtained the best balance between the *P* Scores and *N* Scores for the approach.

We observe that the results for the same configuration with the new approach are particularly better, especially considering the *N* Score and the *ED* class.

Results obtained for task 2 are displayed in Table 5.10. Embeddings generated through our method obtained better results in comparison to the baselines as for *P* the scores are higher, whereas for *N* the scores are lower. For *Baseline 1*, corresponding to our prior method (Section 5.3.2), we present the configuration: $\beta_P = 1$ and $\beta_N = 1$ as it obtained the best *P* Score while keeping a good balance with the *N* Score. Again, better results are obtained by the new approach.

Table 5.9. Task 1 (multi-class) – average cosine similarity evaluation results.

Values for β_P and β_N	P Scores				N Scores			
	ALC	**DEP**	**ED**	**SUI**	**ALC**	**DEP**	**ED**	**SUI**
$\beta_P = 10$ and $\beta_N = 10$ (Prior method – Baseline 1)	0.88	0.85	0.94	0.93	0.37	0.51	0.25	0.53
$\beta_P = 0$ and $\beta_N = 0$ (*Word2vec* – Baseline 2)	**0.96**	**0.96**	**0.96**	0.94	0.96	0.93	0.95	0.75
$\beta_P = 0.01$ and $\beta_N = 0.01$	**0.96**	**0.96**	**0.96**	0.95	0.95	0.92	0.95	0.74
$\beta_P = 0.05$ and $\beta_N = 0.05$	0.92	0.84	0.94	0.81	0.8	0.78	0.59	0.55
$\beta_P = 0.1$ and $\beta_N = 0.1$	0.91	0.83	0.95	**0.96**	0.59	0.66	0.48	0.53
$\beta_P = 0.5$ and $\beta_N = 0.5$	0.9	0.88	0.94	0.95	0.41	0.52	0.19	0.57
$\beta_P = 1$ and $\beta_N = 1$	0.89	0.89	0.94	**0.96**	0.39	0.51	0.18	0.52
$\beta_P = 5$ and $\beta_N = 5$	0.87	0.86	0.94	0.95	0.39	0.51	0.16	0.54
$\beta_P = 10$ and $\beta_N = 10$	0.88	0.86	0.94	0.94	0.34	0.48	0.1	0.52
$\beta_P = 25$ and $\beta_N = 25$	0.86	0.82	0.94	0.93	0.25	0.39	0.04	0.45
$\beta_P = 35$ and $\beta_N = 35$	0.84	0.79	0.93	0.93	0.2	0.34	0.00	0.39
$\beta_P = 50$ and $\beta_N = 50$	0.82	0.76	0.93	0.93	**0.15**	**0.3**	**-0.04**	**0.33**

Table 5.10. Task 2 (binary) – average cosine similarity evaluation results.

Values for β_P and β_N	P Score	N Score
$\beta_P = 1$ and $\beta_N = 1$ (prior method – Baseline 1)	0.68	0.44
$\beta_P = 0$ and $\beta_N = 0$ (*Word2vec* – Baseline 2)	0.66	0.41
$\beta_P = 0.01$ and $\beta_N = 0.01$	0.72	0.41
$\beta_P = 0.05$ and $\beta_N = 0.05$	0.73	0.42
$\beta_P = 0.1$ and $\beta_N = 0.1$	0.74	0.43
$\beta_P = 0.5$ and $\beta_N = 0.5$	**0.75**	0.42
$\beta_P = 1$ and $\beta_N = 1$	**0.75**	0.42
$\beta_P = 5$ and $\beta_N = 5$	0.74	0.42
$\beta_P = 10$ and $\beta_N = 10$	0.74	0.42
$\beta_P = 25$ and $\beta_N = 25$	0.74	0.41
$\beta_P = 35$ and $\beta_N = 35$	0.73	0.40
$\beta_P = 50$ and $\beta_N = 50$	0.73	**0.37**

- Evaluation approaches – visualization

This second evaluation approach allows us to visually observe how some of the predictive terms for each class (top 15 terms according to the X^2 score) are distributed in the vector space without applying our embeddings' generation method (*Word2vec - baseline*), and how they are distributed after its application (*Embedding model 0*).

We also consider the enhanced representation provided by *Embedding model 2* [51].

To generate the plots, Principal Component Analysis (PCA) is used as the dimensionality reduction method to reduce the vectors' dimensions from 100 to 2.

For each plot we report PCA's Total Explained Variance Percentage (TEVP), which is an indicator of the percentage of information retained by the two resulting components, and that is given by the aggregation of the Explained Variance Ratio of each component.

The high rates reported confirm the global quality of the representation. For each task, we retain the configuration which

led to the best results according to the average cosine similarity evaluation.

The results of this evaluation approach for task 1, obtained with $\beta_P = 10$ and $\beta_N = 10$, are displayed in Figure 5.4. This shows that better representations are obtained by *Embedding model 0* and *Embedding model 2* [51], where the vectors of predictive terms from any given class are represented close to each other, while making themselves distinguishable from the vectors of predictive terms for other classes.

It can be seen that suicide and depression related terms cannot be easily separated, which is consistent with the fact that both of these conditions tend to be closely related [146].

Figure 5.5 shows the results for task 2. We consider the $\beta_P = 1$ and $\beta_N = 1$ configuration as the best one according to the *P* score, and it also obtains a reduction in the *N* score in comparison to *Baseline 1* according to the average cosine similarity evaluation. As for task 1, we can notice that the vectors of terms that are predictive for the main class (MEN) are placed closer to the pivot term and thus far from the terms that are predictive for the Control (CON) class with the proposed models.

- Evaluation approaches – predictive task

To evaluate the performance of our embeddings generation approach, we process task 1 (dataset 4a-multiple) and task 2 (dataset 4b-mental).

We define a set of baselines based on state of the art approaches. The same training and test sets exploited in the embeddings generation process are used for this evaluation.

For both tasks considered for evaluation, we define 9 baselines: *Baseline 0* corresponds to a BoW model based on term level {1-2}-grams. More than a baseline, this is a model kept as a reference as we are mainly focused on the evaluation of the models that make use of word embeddings on predictive tasks with small corpora.

Baseline 1, kept as a reference model as well, consists of a model based on features extracted using the lexicons described in Section 4.2.2. These correspond to linguistic dimensions, affective processes and emotions, personal concerns, vocabulary related to risk factors, and topics of interest linked to each condition.

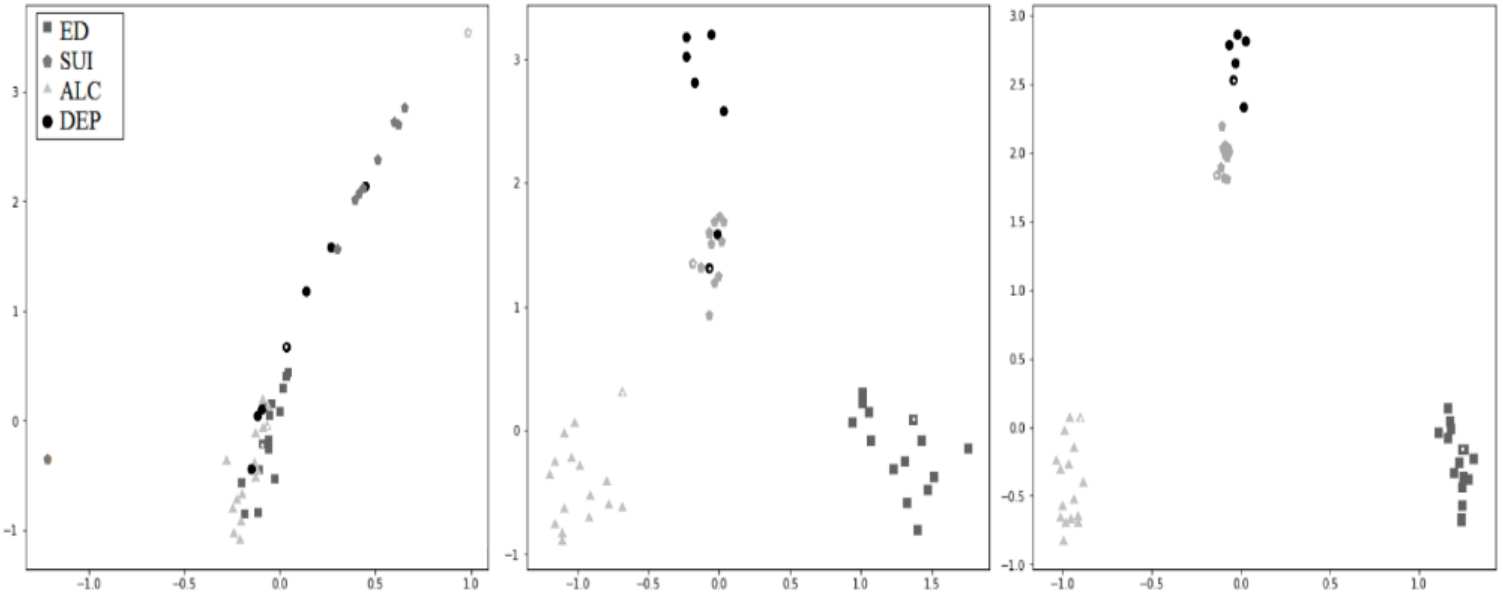

Figure 5.4. Task 1 - Vectors in two dimensions of the top 15 predictive terms of each class. The representations correspond to the 1) Word2vec baseline model (TEVP =97%), 2) Embedding model 0 (TEVP=72%), and 3) Embedding model 2 (TEVP=91%). White dots are placed over pivot terms.

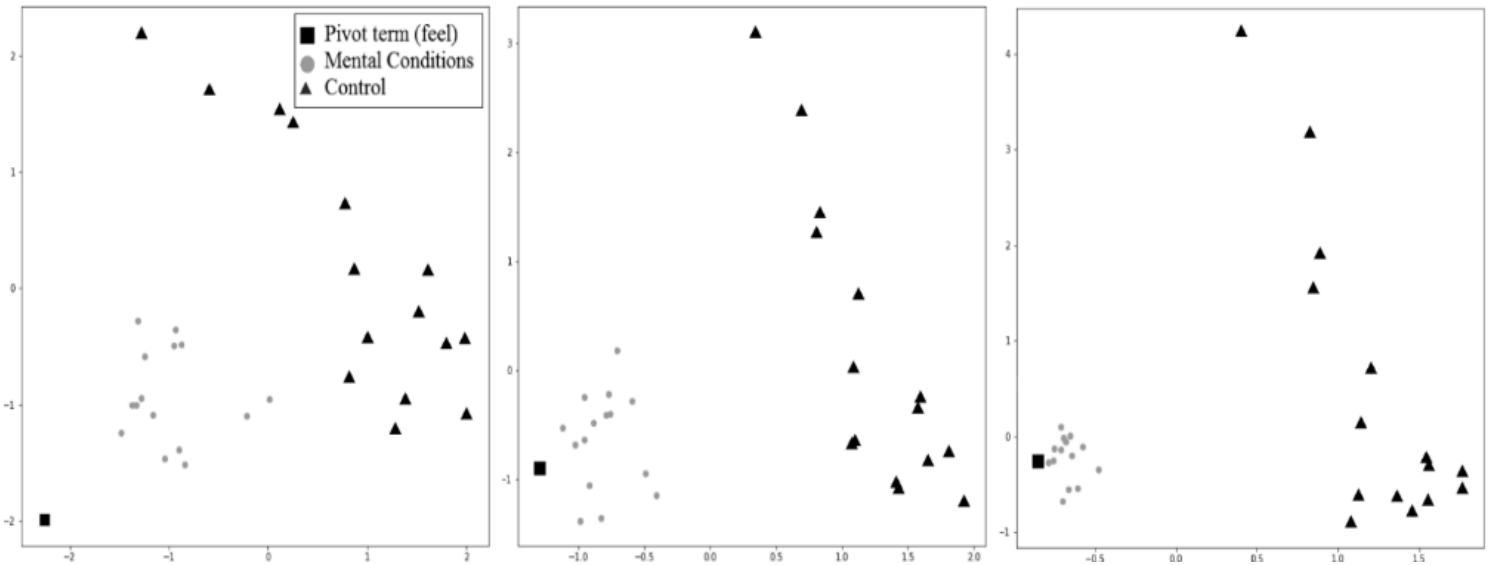

Figure 5.5. Task 2- Vectors in two dimensions of the top 15 predictive terms of each class. The representations correspond to the 1) Word2vec baseline model (TEVP=36%), 2) Embedding model 0 (TEVP=48%), and 3) Embedding model 2 (TEVP=60%). Squares represent pivot terms.

Baseline 2 corresponds to a model that uses DistilBERT context aware pre-trained embeddings with the goal of building a deep learning model with transfer learning. *Baseline 3* consists of using GloVe's pre-trained embeddings without any fine-tuning approach on the domain corpus. *Baseline 4* corresponds to a model where the word embeddings are learned on the training set using the classic Word2vec approach. *Baseline 5* is given by an enhanced version of *Baseline's 4* embeddings, using Faruqui's retrofitting method. *Baseline 6* applies the retrofitting method over GloVe's pre-learned embeddings, while *Baseline 7* corresponds to an embedding model where GloVe's pre-learned embeddings provide the starting weights for learning embeddings on the training set with Word2vec. Finally, *Baseline 8* is a model that uses the embeddings generated using our prior learning approach presented in Section 5.3.2. Table 5.11 shows the baseline models and the proposed embedding variations that they can be compared to.

Table 5.11. Baselines and proposed embedding models (variations) to compare.

Baselines	**Embedding models (variations)**
Baseline 0 (BoW)	All
Baseline 1 (lexicon)	All
Baseline 2 (distilBERT)	All
Baseline 3 (GloVe)	All
Baseline 4 (Word2vec)	Embedding Model 0 (predictive terms)
Baseline 5 (Word2vec + retrofitting)	Embedding model 2 (predictive terms + retrofitting)
Baseline 6 (GloVe + retrofitting)	Embedding model 4 (GloVe's initial weights + predictive terms + retrofitting)
Baseline 7 (GloVe's initial weights + Word2vec)	Embedding model 1 (GloVe's initial weights + predictive terms)
Baseline 8 (prior approach predictive terms)	Embedding model 0 (predictive terms)

We use as classifiers the Scikit Learn Python library implementations for Logistic regression (LR) and Random Forest (RF). These classifiers are trained using a parameter grid search, with a 5-fold cross validation performed for each parameter combination. The model with the best results is kept for its evaluation later over the test set.

We also consider the CNN model previously used in Section 5.2.2. This is denoted as our first deep learning approach DL1. In order to train this model, posts were represented as sequences of terms, and these terms were represented by word embeddings. For the CNN, the embeddings sequences' instances were given as the model input. We used a filter window ({2,3,5} terms). We then applied max pooling and passed the output to either a SoftMax (multi-class task) or Sigmoid (binary task) layer to generate the final output. For Baseline 2, DistilBERT"s output is computed into a single vector with Average Pooling, and later two Dense layers are added to predict the probability of each class; the classifier thus obtained is denoted DL2.

For the deep learning models, 75% of the training instances (posts) were selected for training the model and 25% were considered for validation. Notice that for presenting the results of the deep learning models, we average the results obtained by 5 runs over the test sets (with unseen cases).

For all the classifiers, we defined class weights' parameters for addressing the reduced amount of training samples for certain classes. This was done such a way that all the classes were considered equally important.

For embeddings-based inputs, each instance is represented by an individual post (document) to which a class is assigned. For *Baseline 0* a TF-IDF vectorization of the documents has been applied, considering a list of stop-words and the removal of the n-grams that appeared in less than 20 documents. For *Baseline 1*, we considered as features all the scores obtained for the lexicon categories of Section 4.2.2. For this baseline, each of the categories in Tables 4.1, 4.2, 4.3, and 4.4 were considered as features, along with the 200 prebuilt categories of the Empath tool. To get the score for a category (feature), we consider the frequency of terms belonging to it, then the frequency is normalized by the size (in number of terms) of the full post.

Later, we consider approaches for using embeddings as inputs depending on the classification method selected. The first input, named *aggregation input*, as for the evaluation previously described in Section 5.3.2, it is used for testing machine learning approaches, such as Logistic Regression and Random Forest. It consisted in representing a document through the aggregation of the vector representations of the terms in the document, normalized by the size (words count) of the document. Within this same method, a L2 normalization was applied to all the instances.

The other input approaches were suitable for generating deep learning models, which require the input data to be integer encoded, so that each term is represented by a unique integer, we denote this input as the *Emb. sequence* input. Notice that distilBERT''s input uses a different tokenization approach for which a proper input structure should be provided.

We also build models that use features created through the approach as defined in Equation 5.10. For task 1, a predictive model with 4 features, one per class, was built with this method; each feature corresponds to the PSim value between the document and the pivot term of a class. For task 2, a model with only one feature was built as there is only one pivot term belonging to the main class to predict. These features are referred to as the *PSim* input in our results section.

For both tasks our evaluation measures are: Precision (P), Recall (R), F1-Score (F1) and Accuracy (A). The results for task 1 (multi-class) correspond to the macro average scores for P, R and F1, while their micro average scores are equivalent to the Accuracy. The results for task 2 (binary) for P, R and F1 correspond to the main class to predict (MEN), while we take into account the Accuracy to evaluate the performance of the models for both classes. All the evaluation results correspond to those obtained over the test sets defined for each task, which correspond to cases that have not been seen before by the models, nor have they been used for tuning parameters.

For both tasks, we report the best results obtained for each embedding model, including the baselines. We also report those embeddings models that obtained the best results for each input approach (PSim, Aggregation and Embeddings sequence); and in order to exclusively compare the embeddings models regardless of the input approach, we also present the results of

a single input method (aggregation input) for all the embeddings. This last input approach also corresponds to the method used in Section 5.3.2.

Regarding the parameters of the LR models, for both tasks' models we used a one vs. rest approach with Scikit Learn's liblinear solver. The values of the *C* parameter are defined through a grid search, and its value for each model is mentioned next to the classifier type in each results table.

For the RF classifiers we use Scikit Learn's default parameters except for the number of trees in the forest (n_estimators).

For task 1, according to the results presented in Table 5.12, the type of classification method that obtains the best results for the embedding based inputs is the deep learning model DL1, which obtains the best results for 7 models. Notice that the BoW reference model obtains the best results for the task, which is consistent with the findings in related work addressing similar tasks for the detection of mental health issues [93]. Regardless of this, considering exclusively the approaches based on word embeddings, we observe that the *Embedding Model 4* is the one that obtains the best results for recall, F1-score, and accuracy. Moreover, we can see better results (F1) when the enhanced embeddings models (that use our learning approach) are compared with embedding models that use Word2vec's learning approach (Baselines 4,5 and 7), meaning that $\beta_P = 0$ and $\beta_N = 0$. This can be seen when comparing *Baseline 4* (F1= 65.23%) vs *Embedding Model 0* (F1 = 79.20%); *Baseline 5* (F1 = 68.70%) vs. *Embedding model 2* (F1 = 77.42%); and *Baseline 7* (F1 = 84.31%) vs. *Embedding model 1* (F1 = 86%).

Remember that we consider the Recall and F1-Scores as our most relevant measures. Furthermore, the accuracy of the system cannot be very reliable given the limited number of instances existing for the alcoholism and eating disorders classes.

For the embeddings baselines, the best results (F1) were obtained by *Baseline 6*, which corresponds to the CNN model that considers a retrofitted version of pre-trained GloVe embeddings.

Table 5.12. Task 1 (multi-class) – predictive task evaluation results.

Result type	Models	Input approach	Classifier	P	R	F1	A
Reference baselines	Baseline 0 (BoW)	BoW	LR (C = 10)	**93.66%**	**89.49%**	**91.47%**	**92.48%**
	Baseline 1 (lexicon)	Lexicon scores	LR (C = 100)	37.12%	37.98%	37.27%	63.43%
Embeddings Baselines	Baseline 2 (distilBERT)	Embeddings sequence	DL2	54.00%	69.00%	58.00%	71.00%
	Baseline 3 (GloVe)	Embeddings sequence	DL1	86.67%	83.06%	84.66%	87.17%
	Baseline 4 (Word2vec)	Aggregation	LR (C = 100)	70.11%	62.98%	65.23%	81.77%
	Baseline 5 (Word2vec + retrofitting)	Aggregation	LR (C = 100)	72.27%	66.25%	68.70%	82.27%
	Baseline 6 (GloVe + retrofitting)	Embeddings sequence	DL1	87.65%	83.16%	85.12%	87.84%
	Baseline 7 (GloVe's initial weights + Word2vec)	Embeddings sequence	DL1	**88.82%**	81.26%	84.31%	88.03%
	Baseline 8 (prior approach predictive terms)	Aggregation	DL1	84.09%	75.42%	78.93%	84.67%
Best results for the enhanced embeddings models	Embedding model 0 (predictive terms)	Aggregation	LR (C = 100)	79.69%	78.73%	79.20%	84.29%
	Embedding model 1 (GloVe's initial weights + predictive terms)	Embeddings sequence	DL1	87.97%	84.49%	86.00%	87.38%
	Embedding model 2 (predictive terms + retrofitting)	Aggregation	LR (C = 100)	79.67%	75.58%	77.42%	83.64%
	Embedding model 3 (SVD combination)	Embeddings sequence	DL1	87.78%	82.49%	84.74%	87.24%
	Embedding model 4 (GloVe's initial weights + predictive terms + retrofitting)	Embeddings sequence	DL1	87.74%	**84.64%**	**86.03%**	**88.56%**
Best results for each input approach	Embedding model 4 (GloVe's initial weights + predictive terms + retrofitting)	PSim	LR (C = 5)	76.39%	72.09%	73.99%	80.55%
	Embedding model 1 (GloVe's initial weights)	Aggregation	LR (C = 100)	83.12%	81.06%	82.03%	86.19%
	Embedding model 4 (GloVe's initial weights + predictive terms + retrofitting)	Embeddings sequence	DL1	87.74%	84.64%	86.03%	88.56%

Best baselines and embedding models' variations results for Precision (P), Recall (R), F1-Score (F1) and Accuracy (A). The best results obtained by the configurations are in bold. For all the enhanced embedding models and Baseline 8: $\beta \mathrm{P} = 10$ and $\beta N = 10$.

Regarding the enhanced embedding models, we can observe that embeddings learned exclusively through our approach (*Embedding model 0*) provide better results (F1) compared to Baselines 1, 2, 4, 5 and 8. *Embedding model 4* is also the best model for generating the *PSim* input. This is a promising result for this approach as with only 4 features the Accuracy achieved is only 8.01% lower than the one of the best model (*Embedding model 4* – Embeddings sequence input), and only 11,93% lower than the BoW model.

Table 5.13 shows the results obtained by a single input type (Aggregation input) and classification method (LR) for task 1. Among the embedding models, we observe that the best results in Precision, Recall, F1 Score and Accuracy are given by the *Embedding model 1*. Based on the F1 score, we can see that the embedding models 1 and 4, enhanced through our approach, outperform all the embeddings baselines. Notably, we can observe a 13.97% increase in the F1 Score when comparing the embeddings learned through our approach (*Embedding model 0*) vs. *Baseline 4* (Word2vec). Moreover, we have proved the usefulness of the predictive terms defined through our method for their usage on similar approaches, such as Faruqui's *et al.* retrofitting method [51], where their usage as semantically related terms implied obtaining better results for *Baseline 5* (F1 = 68.70%) vs. *Baseline 4* (F1 = 65.23%); and for *Baseline 6* (F1 = 80.07%) vs. *Baseline 3* (F1 = 79.63%). Also, considering our learning approach combined with the retrofitting method, better results were obtained for the *Embedding model 2* (F1 = 77.42%) vs. *Baseline 5* (F1 = 68.70%) and, the *Embedding model 4* (F1 = 80.50%) vs *Baseline 6* (F1 = 80.07%) cases.

For task 2 the results for the predictive task are presented in Table 5.14. We can see that as for the prior task, the BoW reference model obtains the best results. Addressing the embeddings models, which are our main point of interest, we can see that despite being small, there is an improvement in the results obtained by the enhanced embeddings. This is confirmed by the results shown in Table 5.15 where, as for the prior task, a single input approach is used (Aggregation input). According to Table 5.14, we can observe that the baseline with the best result (F1) is Baseline 6.

Table 5.13. Task 1 (multi-class) – predictive task evaluation results – aggregation input for the enhanced embeddings.

Result type	Models	Input approach	Classifier	P	R	F1	A
Reference baselines	Baseline 0 (BoW)	BoW	LR (C=10)	**93.66%**	**89.49%**	**91.47%**	**92.48%**
	Baseline 1 (lexicon)	Lexicon scores	LR (C=100)	37.12%	37.98%	37.27%	63.43%
Embeddings Baselines	Baseline 2 (distilBERT)	Embeddings sequence	DL2	54.00%	69.00%	58.00%	71.00%
	Baseline 3 (GloVe)	Aggregation	LR (C = 100)	80.17%	79.15%	79.63%	85.04%
	Baseline 4 (Word2vec)	Aggregation	LR (C = 100)	70.11%	62.98%	65.23%	81.77%
	Baseline 5 (Word2vec + retrofitting)	Aggregation	LR (C = 100)	72.27%	66.25%	68.70%	82.27%
	Baseline 6 (GloVe + retrofitting)	Aggregation	LR (C = 100)	80.50%	79.70%	80.07%	84.32%
	Baseline 7 (GloVe's initial weights + Word2vec)	Aggregation	LR (C = 100)	79.55%	80.08%	79.71%	85.87%
Best results for each Model	Embedding model 0 (predictive terms)	Aggregation	LR (C = 100)	79.69%	78.73%	79.20%	84.29%
	Embedding model 1 (GloVe's initial weights + predictive terms)	Aggregation	LR (C = 100)	83.12%	**81.06%**	**82.03%**	**86.19%**
	Embedding model 2 (predictive terms + retrofitting)	Aggregation	LR (C = 100)	79.67%	75.58%	77.42%	83.64%
	Embedding model 3 (SVD combination)	Aggregation	LR (C = 100)	**81.16%**	78.36%	79.69%	85.15%
	Embedding model 4 (GloVe's initial weights + predictive terms + retrofitting)	Aggregation	LR (C = 100)	81.07%	79.97%	80.50%	85.69%

Baselines and embedding models' variations results for Precision (P), Recall (R), F1-Score (F1) and Accuracy (A), using the Aggregation input for all the enhanced embeddings models. The best results obtained by the configurations are in bold. For all the enhanced embedding models and Baseline 8: $\beta_P = 10$ and $\beta_N = 10$.

Table 5.14. Task 2 (binary) – predictive task evaluation results.

Result type	Models	Input approach	Classifier	P	R	F1	A
Reference baselines	Baseline 0 (BoW)	BoW	LR (C = 10)	**98.05%**	**97.45%**	**97.75%**	**98.40%**
	Baseline 1 (lexicon)	Lexicon scores	RF (n_estimators = 100)	68.24%	89.54%	77.45%	81.40%
Embeddings Baselines	Baseline 2 (distilBERT)	Embeddings sequence	DL2	90.87%	93.82%	92.32%	94.43%
	Baseline 3 (GloVe)	Embeddings sequence	DL1	95.96%	96.31%	96.12%	97.23%
	Baseline 4 (Word2vec)	Embeddings sequence	DL1	96.82%	94.26%	95.49%	96.83%
	Baseline 5 (Word2vec + retrofitting)	Embeddings sequence	RF (n_estimators = 1000)	94.60%	96.44%	95.51%	96.77%
	Baseline 6 (GloVe + retrofitting)	Embeddings sequence	DL1	96.09%	96.23%	96.15%	97.25%
	Baseline 7 (GloVe's initial weights + Word2vec)	Embeddings sequence	RF (n_estimators = 1000)	96.89%	94.17%	95.51%	96.84%
	Baseline 8 (prior approach predictive terms)	Embeddings sequence	DL1	95.77%	95.97%	95.82%	97.01%
Best results for the enhanced embeddings models	Embedding model 0 (predictive terms)	Embeddings sequence	RF (n_estimators = 1000)	95.85%	96.40%	96.13%	97.23%
	Embedding model 1 (GloVe's initial weights + predictive terms)	Embeddings sequence	DL1	96.77%	96.19%	**96.46%**	**97.48%**
	Embedding model 2 (predictive terms + retrofitting)	Aggregation	RF (n_estimators = 1000)	95.42%	**96.58%**	96.00%	97.13%
	Embedding model 3 (SVD combination)	Embeddings sequence	DL1	96.05%	95.94%	95.96%	97.11%
	Embedding model 4 (GloVe's initial weights + predictive terms + retrofitting)	Embeddings sequence	DL1	**97.37%**	95.41%	96.37%	97.44%
Best results for each input approach	Embedding model 4 (retrofitting + GloVe's initial weights + predictive terms)	PSim	LR (C = 150)	86.93%	94.50%	90.56%	92.97%
	Embedding model 0 (Predictive terms)	Aggregation	RF (n_estimators = 1000)	95.85%	96.40%	96.13%	97.23%
	Embedding model 1 (GloVe's initial weights + predictive terms)	Embeddings sequence	DL1	96.77%	96.19%	96.46%	97.48%

Best baselines and embedding models' variations results for Precision (P), Recall (R), F1-Score (F1) and Accuracy (A). The best results obtained by the configurations are in bold. For all the enhanced embedding models: β_P=1 and β_N=1.

Table 5.15. Task 2 (binary) – predictive task evaluation results – aggregation input for the enhanced embeddings.

Result type	Models	Input approach	Classifier	P	R	F1	A
Reference baselines	Baseline 0 (BoW)	BoW	LR (C = 10)	***98.05%***	**97.45%**	***97.75%***	***98.40%***
	Baseline 1 (lexicon)	Lexicon scores	RF (n_estimators = 100)	68.24%	89.54%	77.45%	81.40%
Embeddings Baselines	Baseline 2 (distilBERT)	Embeddings sequence	DL2	90.87%	93.82%	92.32%	94.43%
	Baseline 3 (GloVe)	Aggregation	LR (C = 100)	90.99%	97.34%	94.06%	95.61%
	Baseline 4 (Word2vec)	Aggregation	RF (n_estimators = 1000)	94.57%	95.90%	95.23%	96.58%
	Baseline 5 (Word2vec + retrofitting)	Aggregation	RF (n_estimators = 1000)	94.60%	96.44%	95.51%	96.77%
	Baseline 6 (GloVe + retrofitting)	Aggregation	RF (n_estimators = 1000)	95.11%	93.64%	94.36%	96.01%
	Baseline 7 (GloVe's initial weights + Word2vec)	Aggregation	RF (n_estimators = 1000)	96.89%	94.17%	95.51%	96.84%
Best results for the enhanced embeddings models	Embedding model 0 (predictive terms)	Aggregation	RF (n_estimators = 1000)	95.85%	96.40%	***96.13%***	***97.23%***
	Embedding model 1 (GloVe's initial weights + predictive terms)	Aggregation	RF (n_estimators = 1000)	**97.23%**	94.61%	95.90%	97.11%
	Embedding model 2 (predictive terms + retrofitting)	Aggregation	RF (n_estimators = 1000)	95.42%	96.58%	96.00%	97.13%
	Embedding model 3 (SVD combination)	Aggregation	LR (C = 100)	91.35%	***97.55%***	94.35%	95.83%
	Embedding model 4 (GloVe's initial weights + predictive terms + retrofitting)	Aggregation	RF (n_estimators = 1000)	96.90%	94.46%	95.67%	96.95%

Baselines and embedding models' variations results for Precision (P), Recall (R), F1-Score (F1) and Accuracy (A), using the Aggregation input for all the enhanced embeddings models. The best results obtained by the configurations are in bold. For all the enhanced embedding models: $\beta_P = 1$ and $\beta_N = 1$.

We can also see that the best results (P, R, F1 and A) for all the embedding models are given by variations of the enhanced embeddings. In addition, the best PSim result (F1) is only 5.9% lower than the best embedding model score (Embedding model 1 – Embeddings sequence input). For this particular task, we can notice that despite obtaining better results with the enhanced embeddings, the differences with the baselines are minimal.

5.4 Early risk detection models

5.4.1 Introduction

Automated methods have been developed in order to detect depression and other mental illnesses by analyzing user-generated data in social media [26,63]. These methods usually rely on classification algorithms that do not consider the delay in detecting positive cases. In this sense, Losada *et al.* [91] proposed a temporal-aware risk detection benchmark in order to not only consider the accuracy of the decisions taken by the algorithms, but also the temporal dimension.

We address tasks dedicated to the early detection of signs of depression and anorexia [93]. We propose models to sequentially process texts posted by users in social media, and detect traces of depression or anorexia as early as possible [98]. The texts are meant to be processed in the order they were created for a further capability of the system to analyze the interaction between users in social networks, in real time.

In this section we describe the tasks addressed, our research proposal focusing on the feature extraction process, and the learning algorithms used for both tasks.

5.4.2 Tasks

We address two tasks: one dedicated to the detection of users with depression (T1) and another dedicated to the detection of users with anorexia nervosa (T2). Both tasks consisted in analyzing Reddit data (Dataset 1a - depression and Dataset 1b – anorexia) composed by chronologically ordered writings (posts or comments) from a set of social media users [93]. For T1, users were labeled as depressed and non-depressed, and for T2, users were labeled as anorexic and non-anorexic. Given that the dataset corresponded to a shared task

[93], the collection of writings of each user was split into 10 chunks, with a 10% of the total stored messages of the user in each chunk. For evaluation purposes, the chunks were defined so that a detection system can emit a decision after seeing a given amount of chunks.

5.4.3 Models

The main objective of our proposal is to detect cases of depression and anorexia as soon as possible, minimizing the time taken to generate a decision and maximizing the F1 Score. We use machine learning techniques that combine a set of features extracted from the concatenated writings of users on social media. With these features, a model is trained to be applied afterwards to process the users' test text streams, for each task's dataset. To process the writings, the *dynamic method* proposed in [91] is used. This method consisted in building incrementally a representation of each user, and then applying a classifier, which was previously trained with all the users' writings. Following this approach, a decision is made if the classifier outputs a confidence value above a given threshold.

a) Features extracted

The features we considered aim to characterize the content of the writings. The details on these features are explained below, and a summary can be found in Table 5.16.

Table 5.16. Features considered for T1 and T2 in the models evaluated.

Feature Type	Details and resources	Task for which the feature was applied
Linguistic and psychological Processes, and depression Vocabulary	LIWC	T1 and T2
	depression vocabulary	T1
	anorexia vocabulary	T2
N-grams	unigrams	T1 and T2
	bigrams	T1 and T2
Features with added weighted scores	addition of the weighted scores of depression related features	T1
	addition of the weighted scores of anorexia related features	T2

- Linguistic, psychological processes and depression related vocabulary

These are features given by the frequency of words belonging to the categories of the LIWC dictionary [101]. Scores based on linguistic and psychological processes, as well as personal concerns and spoken categories were obtained. They were calculated by normalizing the frequencies of words by the total number of words in the writings of a user.

For T1, two additional domain-based features were obtained by defining antidepressants and absolutist words categories. In this sense, a list of the 10 leading psychiatric drugs as published in [105], and a set of absolutist words based on the work of [5] were added. This last study concluded that the elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation.

For T2, in addition to the LIWC features, 9 features were defined by creating categories of words that belonged to domains related with AN. The vocabulary for these categories was obtained from the codebook's domains and sample keywords defined in [11]. The domains are: anorexia, body image, body weight, food and meals, eating, caloric restriction, binge, compensatory behavior, and exercises. Each domain was defined by a list of keywords as stated in [11].

- N-grams

For the implemented approaches, a TF-IDF vectorization was done from the unigrams, and bigrams of the training set writings. The content of a document was defined by the concatenation of all the writings of a user from all the chunks, in the training phase.

- Feature with weighted scores

For T1, an additional feature was defined by adding the LIWC scores of certain features. they were selected based on the top 4 LIWC categories that were strongly correlated with positive depression cases, as stated in [5]. Antidepressants and absolutist words categories were added too.

In the same way, for T2, a feature was obtained based on the combination of the LIWC score of the 9 features considering the categories of words that belonged to domains related with anorexia [11].

b) Learning Algorithms

Two prediction methods were explored, i.e., Logistic Regression and Random Forest, as they have been used previously as classifiers for similar tasks [91,120].

c) Evaluation measures

For the evaluation of the performance of our methods we report the Precision, Recall and F1-Score. In addition to these commonly known measures, we evaluate our proposal in terms of the ERDE [91], which is a time aware measure that penalizes a delay in the detection of cases of risk. The delay is measured taking into account the amount of writings that the model requires to see before generating a decision. It gives a cost *c* to each binary decision *d* taken by the system at a number *k* of textual items seen before making a decision. This error is defined by Equation 5.11 [91].

$$ERDE_o(d,k) = \begin{cases} c_{fp} & if\ d = False\ Positive\ (FP) \\ c_{fn} & if\ d = False\ Negative\ (FN) \\ lc_o(k).c_{tp} & if\ d = True\ Positive\ (TP) \\ 0 & if\ d = True\ Negative\ (TN) \end{cases} \quad (5.11)$$

As the setting of c_{fp} and c_{fn} depends on the application domain and the implications of FP and FN decisions we adopt the values assigned in [93] for all the costs. The value of c_{fn}=1, while c_{fp} was set according to the proportion of positive cases in each dataset. Given that a late detection should be penalized through this evaluation approach we set $c_{tp} = c_{fn}$. The factor $lc_o(k)$ ($\in$ [0, 1]) represents a cost associated with the delay in detecting true positives, as defined by equation 5.12, where *o* is a parameter that defines the point at which the cost grows more quickly. The overall error is the mean of the ERDE values of all the instances evaluated (users).

$$lc_o = 1 - \frac{1}{1 + e^{k-o}} \quad (5.12)$$

5.4.4 Experimental setup

Using the training data provided for T1, we applied 10-fold cross validation and optimized the parameters through grid search in order to maximize the F1 Score. Each instance of this dataset was defined by the features described in Table 5.16 and represented one user. For each user, the features were extracted from the sequentially-concatenated writings of all their

chunks. The provided test set allowed us to evaluate the behavior of the dynamic method. Also, a part of the training set was used to define a threshold that represents the minimum probability value required by an instance to be classified as positive. The definition of this threshold contributed to the minimization of the ERDE. The performance of the method was evaluated in terms of the evaluation measures.

Similarly for T2, with part of the training data provided we chose to do a 10-fold cross validation combined with grid search in order to optimize the parameters of the algorithms used. The models obtained were afterwards used to process the writings of the test data, applying the dynamic method.

We designed four different models for each task (Table 5.17). Each model contained a particular set of features, and was created by applying either Logistic Regression or Random Forest classifiers.

Table 5.17. Description of the models designed.

Model	T1		T2	
	Features	**Configuration**	**Features**	**Configuration**
Model 1	LIWC: 64 features. Unigrams: 12655 features	Logistic regression Threshold = 0.75	LIWC: 64 features Unigrams: 4303 features	Logistic regression Threshold = 0.75
Model 2	LIWC: 64 features Unigrams: 12655 features	Random forest Threshold = 0.5	LIWC: 64 features Unigrams: 4303 features	Random forest Threshold = 0.5
Model 3	LIWC: 64 features Unigrams and bigrams: 18006 features Depression vocabulary: 2 features Feature with depression weighted scores: 1 feature	Logistic regression Threshold = 0.75	LIWC: 64 features Unigrams and bigrams: 4970 features Anorexia vocabulary: 9 features Feature with anorexia weighted scores: 1 feature	Logistic regression Threshold = 0.75
Model 4	LIWC: 64 features Unigrams and bigrams: 18006 features Depression vocabulary: 2 features Feature with depression weighted scores: 1 feature	Random forest Threshold = 0.55	LIWC: 64 features Unigrams and bigrams: 4970 features Anorexia vocabulary: 9 features Feature with anorexia weighted scores: 1 feature	Random forest Threshold = 0.55

5.4.5 Results

- Task 1: Depression

The best F1 score value (0.55) was provided by model 1, with LR, the use of unigrams, and the LIWC categories. The best ERDE {5,50} scores were reported by the same model. Table 5.18 also reports Precision, Recall and ERDE scores for Model 1.

Regarding ERDE, Table 5.18 reports 4 different ERDE measures, organized in 2 subsets: the first one corresponds to ERDE {5,50} chunks, these are scores calculated at a chunk level meaning that the amount of writings considered is equal to the accumulated number of writings of all the chunks that were seen by the system before emitting a decision, this is the evaluation approach selected by [93] recalling that the models reported were submitted to participate in a shared task [93]. The second is the ERDE {5,50} writings, where the scores are calculated with the exact number of writings that were analyzed before deciding.

The results show that processing the streams dynamically, writing per writing, instead of chunk by chunk, reduces the ERDE value. Also, the Logistic Regression classifiers provided better results compared to the models where Random Forest was applied.

Table 5.18. Top ranked models regarding F1 score (T1 and T2)

Task	Model	F1	P	R	ERDE 5 chunks	ERDE 50 chunks	ERDE 5 writings	ERDE 50 writings
T1	Model 1	0.55	0.56	0.54	9.39%	7.35%	9.11%	6.41%
T2	Model 3	0.73	0.73	0.71	12.19%	9.74%	10.48%	8.17%

Table 5.19 reports the results after processing each chunk. Focusing on T1, as more chunks are analyzed, the F1 score increases, and so the precision and recall. The ERDE decreases after analyzing the second chunk, and starts to slightly increase afterwards. Regarding ERDE50, the percentage mostly decreases after processing a new chunk. With all chunks processed, we found that the system got the highest amount of true positive cases (47%), right after processing the first chunk, but this is precisely when the highest amount of false positive cases are predicted too (76%).

Table 5.19. Results obtained after processing each chunk (T1 and T2).

Task	Measure	Chunk									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
T1	**F1**	0.32	0.40	0.44	0.48	0.51	0.51	0.52	0.51	0.52	**0.55**
	P	0.43	0.49	0.52	0.54	0.55	0.55	0.55	0.53	0.54	**0.56**
	R	0.25	0.34	0.38	0.43	0.47	0.48	0.49	0.49	0.51	**0.54**
	ERDE5	9.26%	**9.04%**	**9.04%**	9.05%	9.06%	9.07%	9.08%	9.11%	9.11%	9.11%
	ERDE50	7.99%	7.16%	7.04%	6.72%	6.73%	6.62%	6.63%	6.65%	6.65%	**6.41%**
T2	**F1**	0.47	0.55	0.60	0.61	0.62	0.64	0.69	0.72	0.72	**0.73**
	P	0.74	0.75	**0.77**	0.75	0.73	0.74	0.76	0.76	0.76	0.76
	R	0.34	0.44	0.49	0.51	0.54	0.56	0.63	0.68	0.68	**0.71**
	ERDE5	10.92%	**10.36%**	10.36%	10.40%	10.44%	10.44%	10.44%	10.48%	10.48%	10.48%
	ERDE50	9.26%	**8.68%**	8.68%	8.72%	8.45%	8.45%	8.13%	8.17%	8.17%	8.17%

- Task 2: Anorexia

The best model for T2 was Model 3 with a F1 Score of 0.73. Regarding ERDE score, Model 4 reported the best score for ERDE5 (12.93%) and Model 1 for ERDE50 (11.34%). As in T1, Table 5.18 displays the ERDE chunks and ERDE writings. We can see that processing the streams writing per writing and using Logistic Regression classifiers provided better results.

From Table 5.19 we observe that, even though the recall increases considerably after processing each chunk, the precision seems to remain stable. The ERDE percentages seem to present a similar pattern as for T1. After processing chunk 1, the highest amount of true positives are detected (48%), and again the highest amount of false positive cases are identified (56%).

In [129] we extended the work addressing anorexia nervosa integrating topic modeling to the models evaluated. The best model of that work obtained a F1 score of 0.85 with an ERDE5 of 13.05% and an ERDE50 of 7.26%.

5.5 Analysis of biases

5.5.1 Introduction

Algorithmic bias is defined as a "systematic deviation in an algorithm output, performance, or impact, relative to some norm or standard" [53]. Walsh *et al.* [162] state that health disparities contribute to algorithmic bias. These can be cultural dissimilarities, differences in the relation between patients and clinicians with different backgrounds, or prevailing societal notions about the susceptibility of certain groups to mental illness. An instance of this is the notion of women having a higher prevalence of depression. These notions can incorporate bias in underlying data and model specifications. Consequently, they can influence the reliability of predictive models for their actual deployment in real-life settings [162].

In this section we describe our contribution to a work dedicated to analyze gender bias in models for the detection of anorexia [149]. These models were generated using selected instances of dataset 3 – anorexia nervosa.

Our main contribution to this work is the analysis of insights regarding the input data used to generate predictive models that were found to exhibit relevant biases in the false negative rate

(proportion of positives which yield negative test outcomes with the test, FNR). The FNR was higher for females in comparison to male cases despite having more instances of female users in the dataset. In fact, the overall performance (F1-Score) obtained for males (F1=0.95) was significantly higher compared to females (F1=0.84) [149]. The analysis of this aspect is relevant as a false negative in this context can lead to the omission of proper treatment for people at risk.

5.5.2 Dataset instances analyzed

We analyzed instances of Dataset 3 defining 2 groups for assessment: 1) anorexia nervosa cases, which consisted in a total of 177 users that were part of the anorexia and treatment groups; and 2) control cases (326 users) consisting of instances of the focused and random control groups of dataset 3.

We only considered instances with all the features values complete and discarded control instances that correspond to organizations. Table 5.20 describes the instances considered for each group.

Table 5.20. Description of the instances analyzed from Dataset 3 - anorexia nervosa.

Description	Positive (AN)	Control
No. samples	177	326
Female	127	157
Male	50	169

Taking into account the features extracted for their analysis in Section 4.4.2, we define groups of features for bias characterization as described in Table 5.21.

5.5.3 Bias characterization

With the purpose of investigating the causes of the algorithmic bias when assessing AN on social media, we studied the features considered as input for the predictive models to identify which of those variables (see Table 5.21) are more predictive for each gender.

Table 5.21. Types of features explored for bias analysis.

Types of features	Description
Content shared and interests	Linguistic dimensions Affective processes and emotions Personal concerns Risk factors vocabulary Anorexia related vocabulary Topics of interest Proportion of anorexia nervosa related tweets
Social network	Measures of interactions and engagement Analysis of followees and communities detection Analysis of interests between users and their followees
Behavioral aspects	Activity on a daily, weekly and monthly basis Sleep period tweeting ratio
Demographics	Gender Age

We separated the instances by gender, and proceeded to apply feature selection approaches. In particular, we considered Recursive Feature Elimination (RFE) [167] in order to analyze the relevance of features depending on the gender of the users.

RFE starts with all features and then a subset of k features (the most relevant) is searched by removing features until the desired number remains. It works by training an estimator on the initial set of features; then, features are ranked by importance based on the estimator.

Afterwards, features that are less important are removed sequentially from the current set of features so that the process can be recursively repeated on the pruned set until the number k of desired features to keep is reached.

For our case, we used a Logistic Regression estimator, and obtained a rank for all the features used by assigning the value of 1 to k, as it provides a rank based on the order in which features were removed at each iteration until only one feature is left. We used Python's Sklearn RFE feature selection implementation [115]. Considering the top-10 (i.e., k=10) features selected through this approach for each gender model, we make comparative plots of their distributions in order to observe how the values of the selected features differ.

In order to investigate if the models selected the same features as a group of real experts on eating disorders would do, we asked 5 clinicians to answer a survey.

These clinicians were 5 experts that had participated in social media writings' labeling tasks. They were asked to assign a level of importance to the different feature types extracted from the dataset (considering that they should predict AN risk just based on writings, as our models do). These feature types (Table 5.21) explore the usage of grammatical and syntactical elements; the usage of terms related to emotions, personal concerns, social support received, biological processes and health, suicide risk factors, and eating disorders related vocabulary. We also took into account behavioral patterns that imply a prolonged use of social media; and demographic elements such as age and gender.

The importance levels ranged between 1 and 5, where assigning a score of 1 meant that the feature type was not relevant. Whereas a score of 5 meant the feature type was very important for screening anorexia nervosa. Clinicians were allowed to add comments regarding the feature types suggested.

We calculated means, medians, and the standard deviation of the scores assigned to each feature type, and applied different approaches to measure the inter-rater agreement.

Based on the experts' assessment results, we proceeded to compare their feature types' importance with the relevance assigned by a predictive model trained over all the instances, and features.

We use the RFE's rank of the generic model and assign to each feature a score equivalent to its inverse rank position, meaning that the feature ranked first gets a score equivalent to the rank of the last feature in the ranking. This score corresponds to the importance level assigned to the feature based on an automated predictive model.

Later, each feature is mapped to the feature type that it belongs to in order to average the scores obtained by all the features belonging to a given feature type. Once a single score is obtained for every feature type, we proceed to compare the scores obtained by the classifier with the scores assigned by the experts. A normalization process is applied before, in order to

scale the scores of each group (model and experts) between 0 and 1.

Notice that we considered the feature that measures the proportion of anorexia nervosa related tweets as part of the anorexia related vocabulary feature type. This was done because the feature is given by a deep learning classifier that takes as input word embeddings, which are vector representations of the terms found in the users' writings, and users with AN are more likely to make use of such terms.

Table 5.22 shows the top-10 features selected according to the RFE approach for each gender model using a Logistic Regression estimator for both cases. We also show the top-10 features given by a model ("Generic model") with all the instances (males and females) using gender as a feature.

We can see that for all the models the most relevant features measure the usage of first person singular pronouns and the proportion of anorexia nervosa related tweets. "Hate", as a suicide risk factor, and "sadness" are features that are also important for all the models.

The distribution of the top-10 features for the female and male models are displayed in figures 5.6 and 5.7. Notice that the feature that measures the proportion of anorexia nervosa related tweets implies the usage of anorexia related vocabulary.

Figure 5.8 shows a comparison of the importance of each feature type for each gender model. We can notice that eating disorders' related vocabulary is the most relevant for both genders, whereas biological processes and suicide risk factors are the most relevant for males, and age, emotions and personal concerns are relevant for females.

Table 5.23 shows the results of the survey performed to clinicians to know the most important features they consider when assessing AN based on writings. We averaged the relevance scores assigned by the clinicians participating in the survey

Considering each question as a case and our 5 annotators as raters, we use two inter-rater agreement measures suitable for studies with more than two raters: Fleiss Kappa (κ=0.20) [56] and the Intraclass Correlation coefficient (ICC=0.87) [147]. Among these measures, the ICC is one of the most commonly-used statistics for assessing inter-rater reliability for ordinal variables [66].

Table 5.22. Top 10 features selected according to the RFE approach (* = features relevant for both models).

Rank	Female model	Feature Type	Male model	Feature type	Generic model	Feature type
1	First person singular pronouns*	Grammatical and syntactical elements	First person singular pronouns*	Grammatical and syntactical elements	First person singular pronouns*	Grammatical and syntactical elements
2	Tweets' classifier median score*	Proportion of AN related tweets	Tweets' classifier median score*	Proportion of AN related tweets	Tweets' classifier median score*	Proportion of AN related tweets
3	Work	Personal concerns	Anxiety	Affective processes and emotions	Sadness*	Affective processes and emotions
4	Feeling	Affective processes and emotions	Sadness*	Affective processes and emotions	Suicide risk factors: hate*	Suicide risk factors
5	Suicide risk factors: hate*	Suicide risk factors	Suicide risk factors: hate*	Suicide risk factors	Articles	Grammatical and syntactical elements
6	Sadness*	Affective processes and emotions	Articles	Grammatical and syntactical elements	Biological processes	Biological processes and health
7	Exercise	Anorexia related vocabulary	Disgust	Affective processes and emotions	Negative emotions	Affective processes and emotions
8	Biological processes	Biological processes and health	Food and meals	Anorexia related vocabulary	Food and meals	Anorexia related vocabulary
9	First person pronouns (plural)	Grammatical and syntactical elements	Past	Grammatical and syntactical elements	Past	Grammatical and syntactical elements
10	Trust	Affective processes and emotions	Third person pronouns (plural)	Grammatical and syntactical elements	Suicide risk factors: self-loathing	Suicide risk factors

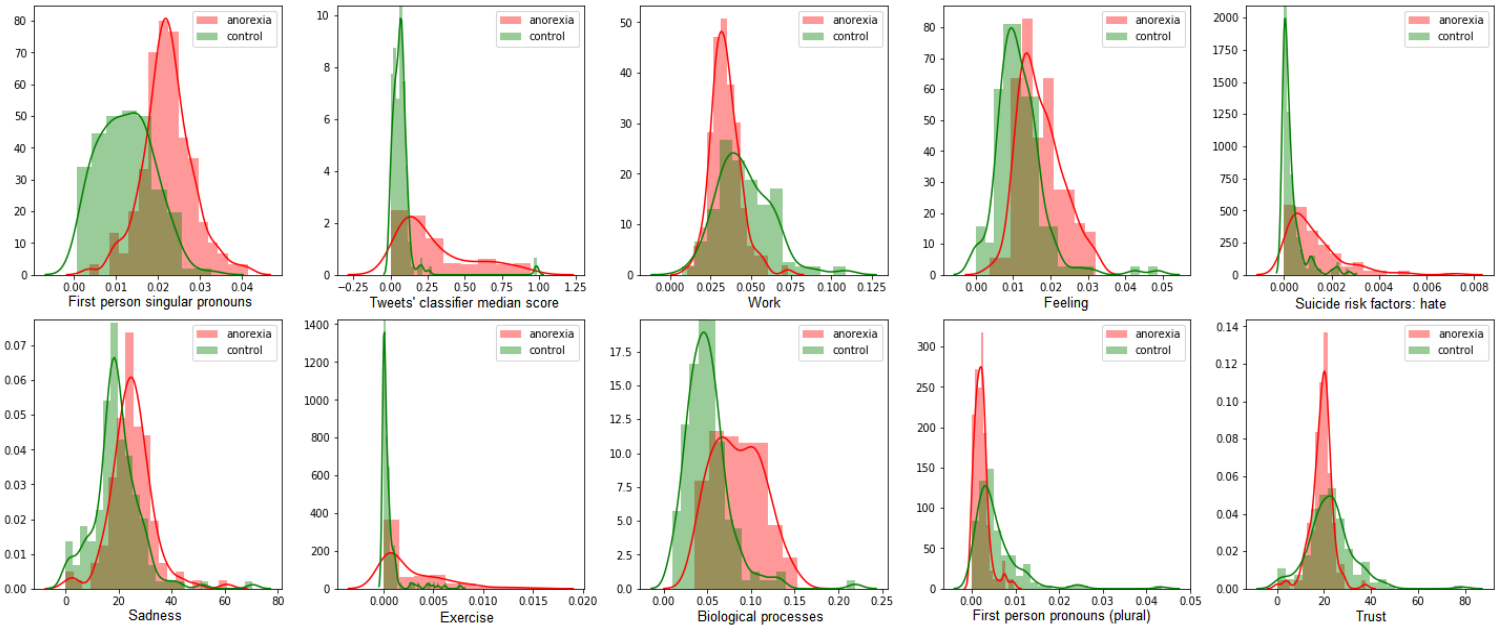

Figure 5.6. Top-10 features selected by the female data model.

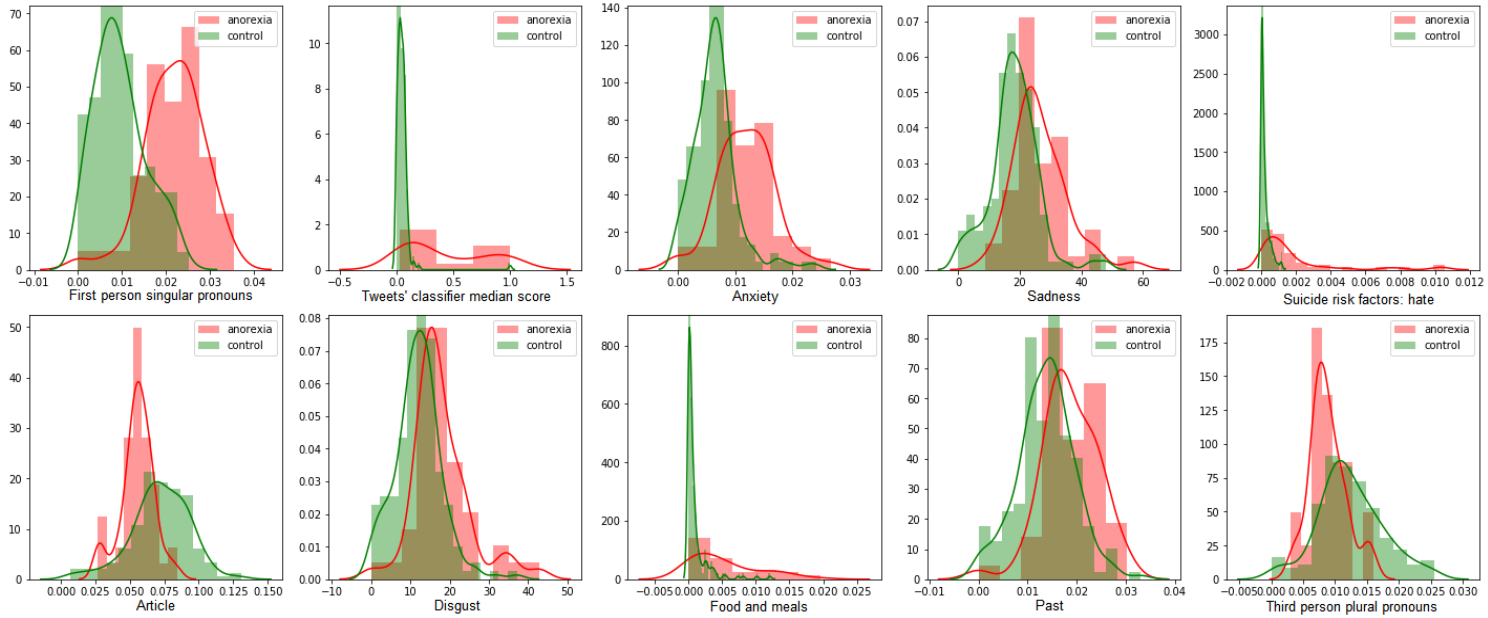

Figure 5.7. Top-10 features selected by the male data model.

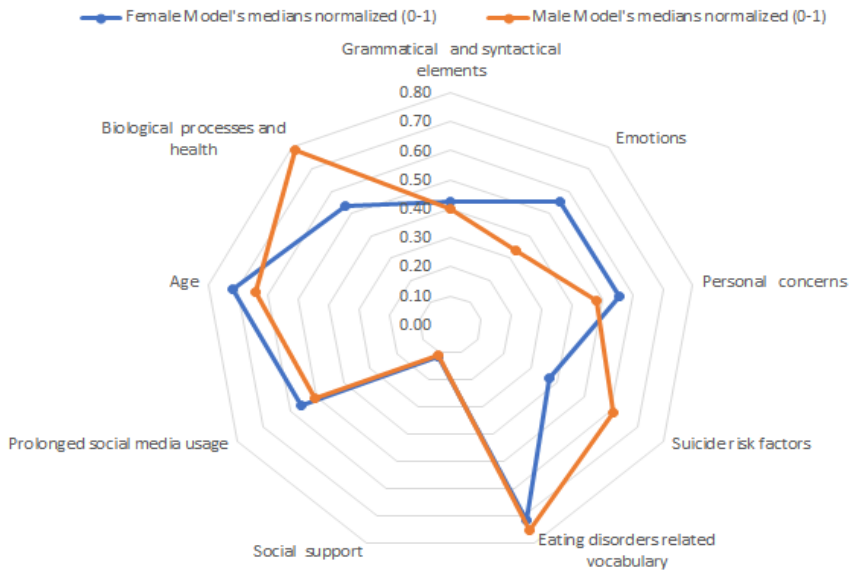

Figure 5.8. Feature importance for each gender model.

The ICC results, which are more suitable for ordinal data suggest a good reliability, whereas κ indicates a slight agreement.

We also calculate the percent agreement [7] for multiple raters, where the individual agreement for each feature type is described in Table 5.23. The average percent agreement is 44%, which implies a moderate agreement.

The feature types that raters found most relevant were the ones that measured the usage of eating disorder's related vocabulary, with a full agreement among clinicians, along with suicide risk factors, biological processes and health, and gender. The least relevant feature type was related with the usage of grammatical and syntactical elements.

The survey also asked for the factors that are taken into account by clinicians in a medical consultation for anorexia nervosa screening. In this case, experts mentioned aspects such as weight, height, restrictive behaviors, obsessive personality, purgative behaviors, body mass index, fear to gain weight, daily life issues (work, school, personal relationships), family members with prior eating disorders, different physical indicators (thermoregulation difficulties and bradycardia), low self-esteem, and gender, as women are more likely to be diagnosed with this type of eating disorder.

Table 5.23. Results on the survey answered by clinicians on the most important features for assessing AN.

Feature type	Description	Average relevance (1-5)	Mode (1-5)	Median (1-5)	Standard deviation	Percent agreement
Grammatical and syntactical elements	Usage of grammatical and syntactic elements, such as personal pronouns, verbs, etc.	1.60	1	1	1.20	0.60
Emotions	Usage of terms related to emotions such as joy, sadness, fear, etc.	3.60	4	4	0.49	0.40
Personal concerns	Usage of terms related with personal concerns such as work, leisure, religion, etc.	3.00	3	3	1.26	0.30
Social support	Usage of terms related to social support as friends, family, loneliness, etc.	3.60	4	4	0.49	0.40
Biological processes and health	Usage of terms related with biological processes and health as eating, therapy, healing, etc.	4.20	5	4	0.75	0.20
Suicide risk factors	Usage of terms related with suicide risk factors as self-harm, bullying, substance abuse, etc.	4.60	5	5	0.49	0.40
Eating disorders related vocabulary	Usage of terms related to eating disorders such as laxative names, weight concerns, etc.	5.00	5	5	0.00	1.00
Prolonged social media usage	Posting frequency.	4.40	4	4	0.49	0.40
Age	User age.	4.00	4	4	0.63	0.30
Gender	User gender.	4.60	5	5	0.49	0.40

When comparing the feature types that are relevant according to the RFE method applied over the generic model, and the ones that are relevant for experts (see Figure 5.9 and Table 5.24), we can observe that for the predictive model the most relevant feature types are the age, eating disorders related vocabulary, and biological processes and health.

Notice that the model and clinicians agree on the fact that eating disorders related vocabulary is relevant, whereas clinicians also assign a high relevance to suicide risk factors and gender. The feature types that the model considers to be less relevant are social support and prolonged social media usage, whereas clinicians assigned grammatical and syntactical elements as the less relevant.

Finally, the fact that suicide risk factors seem to be less relevant for the model is because they are given by lexicons with a limited amount of keywords, which do not necessarily always capture the existence of a given risk factor, as it cannot always be explicitly described in the text. Clinicians, on the other hand, are capable of identifying suicide risk factors that are described implicitly in the text, and handle a wide vocabulary in comparison to the model.

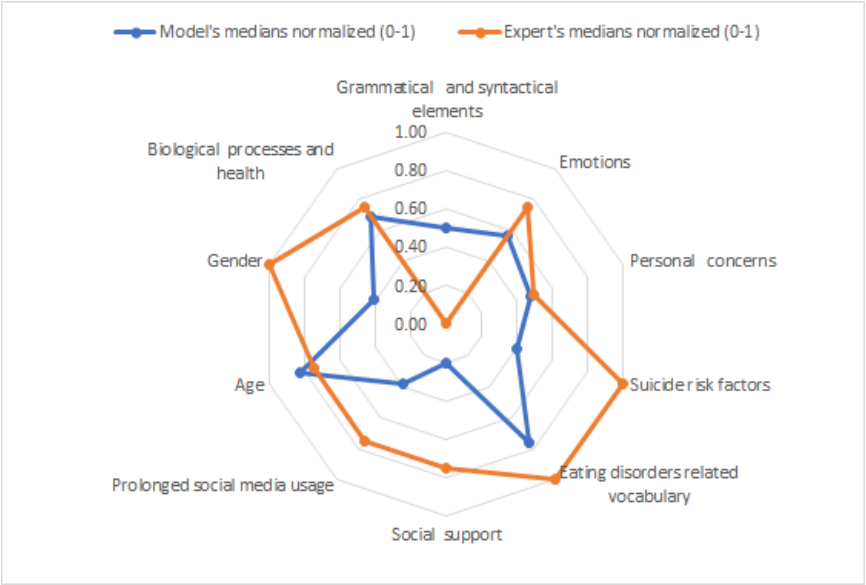

Figure 5.9. Feature types relevant for the generic model vs the ones relevant for the clinicians.

Table 5.24. Model vs Experts feature type rankings.

Feature type	Model's medians normalized (0-1)	Model's feature type ranking	Expert's medians normalized (0-1)	Experts' feature type ranking
Age	0.83	1	0.75	2
Eating disorders related vocabulary	0.76	2	1.00	1
Biological processes and health	0.69	3	0.75	2
Emotions	0.57	4	0.75	2
Grammatical and syntactical elements	0.50	5	0.00	4
Personal concerns	0.48	6	0.50	3
Gender	0.41	7	1.00	1
Suicide risk factors	0.40	8	1.00	1
Prolonged social media usage	0.39	9	0.75	2
Social support	0.20	10	0.75	2

5.6 Discussion

In Chapter 5 we have presented several models addressing predictive tasks for mental health state assessment (RQ2). Among these models we have addressed binary and multiclass detection tasks, we have explored the contribution of multiple feature types, we have taken into account early risk detection settings, and we have also explored biases in predictive models.

First, we presented a methodology for suicide risk assessment on social media. We extracted information from multimodal data to build statistical and deep learning–based predictive models. Our models considered a set of features based on BoW and n-grams, lexicons, relational, statistical, and behavioral information, in addition to image analysis. To the best of our knowledge, this was the first approach that addressed the

combination of all these types of features for suicide risk assessment at the user level. Moreover, we highlighted the usefulness of discarding the noise of writings not related to the topic of study through the definition of a short profile version (SPV), which outperformed the baseline given by the analysis of the full profile of the user, with an increase in accuracy and F1.

We also compared the performance of predictive methods trained on different control groups with the goal of making a more specialized classifier capable of distinguishing users at risk from control cases, even when the discussed topic is similar. Better results were achieved in terms of AUC-ROC when using generic control users instead of users who make use of suicidal vocabulary.

We also highlighted the importance of the interpretability of our features, considering elements that can be understood by clinicians and mapped to their screening practice. The results of our experiments showed that within the types of features analyzed, there were multiple significant features that may lead to the detection of risk situations, the most relevant ones were based on the identification of textual and behavioral elements such as self-references, the number of tweets posted, and the time that passes between each post ($p<.001$).

Text-based features were the most relevant for our models; however, their combination with image-based scores, along with relational and behavioral aspects, allowed us to obtain models that outperformed the results provided by an exclusively text-based model.

Considering the relevance of text cues for the generation of predictive models, we presented an approach for enhancing word embeddings towards a binary classification task on the detection of AN. The method developed extends Word2vec considering positive and negative costs for the objective function of a target term. The costs are added by defining predictive terms for each of the target classes. The combination of the generated embeddings with pre-learned embeddings is also evaluated. Our results show that the enhanced embeddings outperform the results obtained by pre-learned embeddings and embeddings learned through Word2vec regardless of the small size of the corpus.

The findings of this work inspired the evaluation of the method on similar tasks, which can be formalized as document categorization problems, addressing small corpora. In this sense we generated word embeddings adapted to domain specific multiclass classification tasks, and in particular we addressed a task dedicated to the detection of cases of different conditions such as suicidal ideation, depression, eating disorders and alcoholism. To create the word embeddings we used methods dedicated to detect predictive terms for a given class as described in Chapter 4 – Section 4.2.3. We also improve our embeddings generation approach to address binary tasks with the detection of mental health issues vs. control cases (MEN vs. control task).

The proposed predictive models obtain the best results in Recall, F1-Score and Accuracy compared to the embeddings-based baselines for both tasks. Results also demonstrate that word embedding based models are less accurate compared to BoW models for these types of tasks. These findings fit the conclusions of related work dedicated to the detection of depression [93].

Another interesting aspect of this work concerns the performance of the predictive models that use the *PSim* features, which were generated using the enhanced embeddings. Through our proposal, with only 4 features the accuracy achieved by the model is only 8.01% lower than the one of the best embeddings model (embeddings sequence input) for the multi-class task, and only 4.51% lower for the binary task.

Later, we also addressed early risk detection settings. We proposed several models for the early detection of cases of depression and anorexia, by dynamically processing users' text streams. Different machine learning approaches were designed using features extracted from the texts. These features were based on linguistic information, domain-specific vocabulary, and psychological processes. The models generated have a better performance for predicting anorexia, while depression seems to be a task for which the approaches proposed do not perform well.

We have also characterized gender bias in predictive models that address anorexia nervosa detection using Twitter data. We analyzed the most relevant features selected by our models for

assessing female and male users separately, and compared these features with those selected by clinicians when classifying risk of AN just based on the writings of the users. We have found that biological processes and suicide risk factors are the most relevant for detecting AN cases in males, and age, emotions and personal concerns are more relevant for female cases, probably because women are more concerned about their body and personal image (biological processes and health) regardless of having an eating disorder, while men do not tend to address these aspects unless there is an underlying health issue.

Comparing the findings of our research work concerning predictive tasks, we have found that in general, text-based features are the most informative for predictive models. We have also observed that among the use cases addressed, for predictive models dedicated to detect suicidal ideation and anorexia in Twitter, similar results are obtained in terms of the F1 score (suicide with F1=0.86 vs. anorexia with F1=0.84).

Through our models addressing Reddit data for depression and anorexia, we have observed that depression detection is the hardest task to address in an early risk detection context (anorexia with F1=0.73 vs. depression with F1=0.55).

Using the same anorexia Reddit dataset, we have also found that the usage of models created with our enhanced word embeddings for the detection of anorexia have improved the performance of the models obtaining a F1 score of 0.77.

With the work addressing all the use cases (dataset 4), we have also found that distinguishing a mental health condition from other conditions is harder than distinguishing any mental condition from control cases.

Also, with this case in particular we have seen that these models obtain the best results compared to all the tasks addressed before, and it might be given by the fact that here we address texts at a post level, instead of at a user level. This can be supported by the results obtained by the tweet (post) level classifier defined to create the SPV in the suicidal ideation detection task (Section 3.4.1), and also by the tweet level classifier created to assess the proportion of AN related tweets in the AN characterization work (Section 4.4).

Chapter 6

HARMLESS CONTACT RECOMMENDATION

6.1 Introduction

In this chapter we describe and evaluate our proposal to help people with mental disorders. It addresses the usage of contact recommender systems to allow users with mental disorders to approach harmless content and also to follow accounts that promote pro-recovery content.

We use anorexia nervosa as our use case. In social media, prior studies have identified two types of communities related to eating disorders: ED communities and pro-recovery communities [163]. They have found that among these communities the communication is mostly intra-cluster. However, we have found a shift in the interests of users as they move towards treatment, meaning that the exposure to pro-recovery content might not lead to its rejection.

We propose a contact recommendation approach suitable for social platforms where users can establish links with others through a *follow* relation. Twitter is an instance of such platforms where, given a user *u*, the users followed by *u* are referred as *u*'s followees, whereas the users following *u* are referred as *u*'s followers. As it can be seen in Figure 6.1, the objective of a common recommendation model [85] is to rank

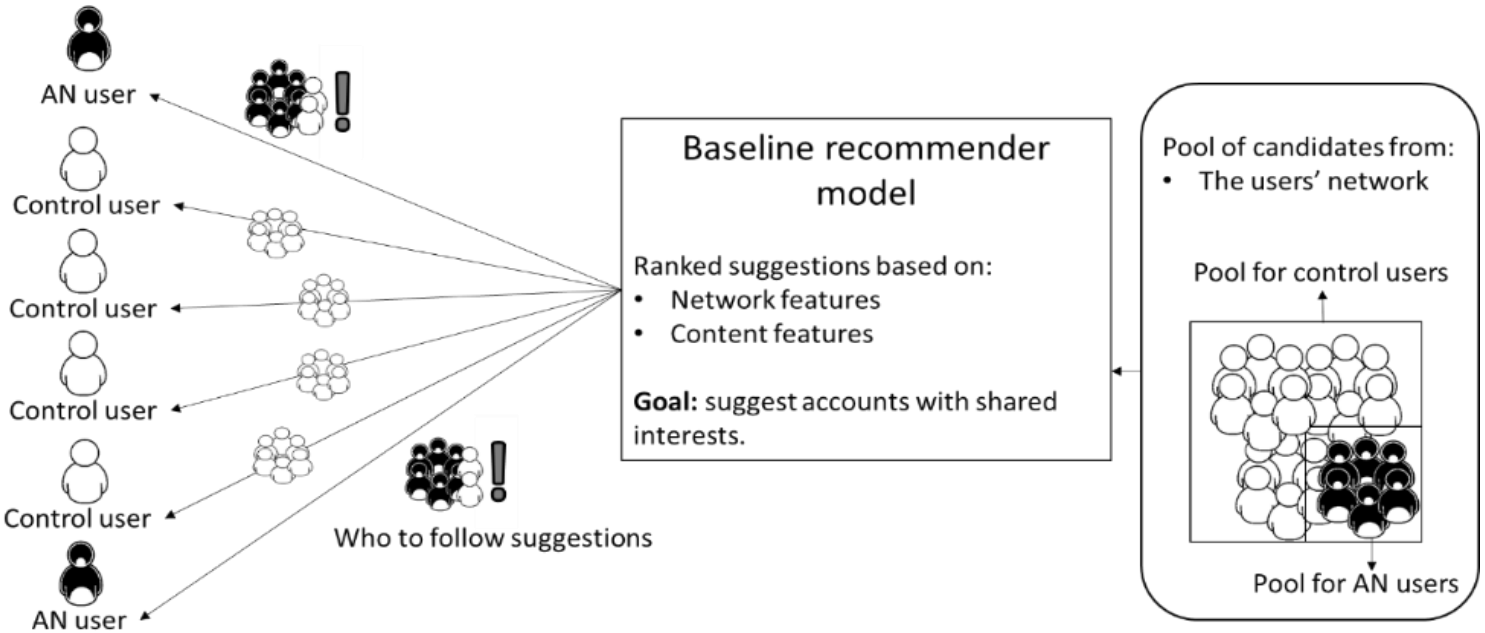

Figure 6.1. Architecture of common contact recommendation models, referred here as a Baseline recommender model, which is potentially harmful for vulnerable users.

on top the accounts that the user is more likely to follow, under the principle that people tend to follow users who they are likely to know (network) or that have interests in common (content). As users with AN are more likely to be following their peers or accounts that promote unhealthy habits (harmful accounts), it is likely for the recommender to provide harmful suggestions as we have observed in Section 4.4.3.

Studies suggest that people surrounded by a support group are likely to recover from mental disorders [27]. However, online ED communities present characteristics of echo chambers and filter bubbles (pro-anorexia), meaning that people that share dissenting opinions (pro-recovery) cannot be reached as they are not likely to be displayed as suggestions [35]. In this sense, we find it relevant to facilitate the communication between people living with eating disorders and pro-recovery communities. This can encourage ED users to seek treatment and to receive support during the recovery process.

The present work seeks to contribute to the development of a contact recommender system for users with AN. We propose an approach, which avoids the recommendation of harmful content to ED users. We do this by recommending accounts with similar yet harmless interests.

Compared to approaches built for detecting and mitigating echo chambers and filter bubbles [31], our goal is not precisely to recommend only pro-recovery accounts (the opposed opinion) but to reduce the number of harmful accounts suggested. These are accounts that not only promote pro-ED content, but also may promote depressive and suicidal thoughts, diets, excessive physical exercise, etc.

It prioritizes the suggestion of harmless content, including pro-recovery accounts, which also share interests with ED users. We believe this is an effective way to favor inter-cluster communication.

Recalling the trans-theoretical model of health behavior change, the contemplation stage is relevant for our work, as this is the stage where people are conscious of an existing issue, yet they simultaneously consider and reject changing their unhealthy habits. This stage is relevant to define our recommendation approach, as users at this stage are more

likely to look for help, which might eventually lead them to reach out for proper treatment.

Through this work we: 1) define a contact recommendation approach for users with anorexia nervosa; 2) we evaluate a classification model to detect users at the contemplation stage; 3) we evaluate a classification model to distinguish harmful from harmless accounts; 4) we propose an evaluation approach that involves the participation of experts, and volunteers with anorexia nervosa; 5) we define a measure that evaluates the performance of the recommendation approach taking into account its precision and the ratio of harmless accounts selected by the user.

6.2 System architecture

Our recommendation approach (Figure 6.2), consists in 1) detecting AN users at the contemplation stage given that the recommendation approach will be applied exclusively over such users; 2) defining a pool of candidates composed by users that are more likely to be harmless. This is done by applying a harmless users' detection model for the definition of the pool of candidates and by introducing a group of pre-labeled pro-recovery users to the pool. Finally, 3) the recommendation model's objective function is defined by a combination of network and content scores with a weight given by a harmlessness factor, which modifies the score of the suggested candidates by penalizing those that are likely to be harmful.

Users are ranked according to the score obtained, and the top K suggestions are displayed to the user. This approach also makes sure that some of the pro-recovery accounts are part of the suggestions displayed.

Depending on the top K suggestions that will be shown to the user, a fixed percentage of these should correspond to those pro-recovery users with the highest scores obtained (based only on the content score).

It is important to mention that the pool of candidates of a target user u is given by their neighborhood as described in [9], meaning that, according to Figure 6.3, it is defined by users from level 3.

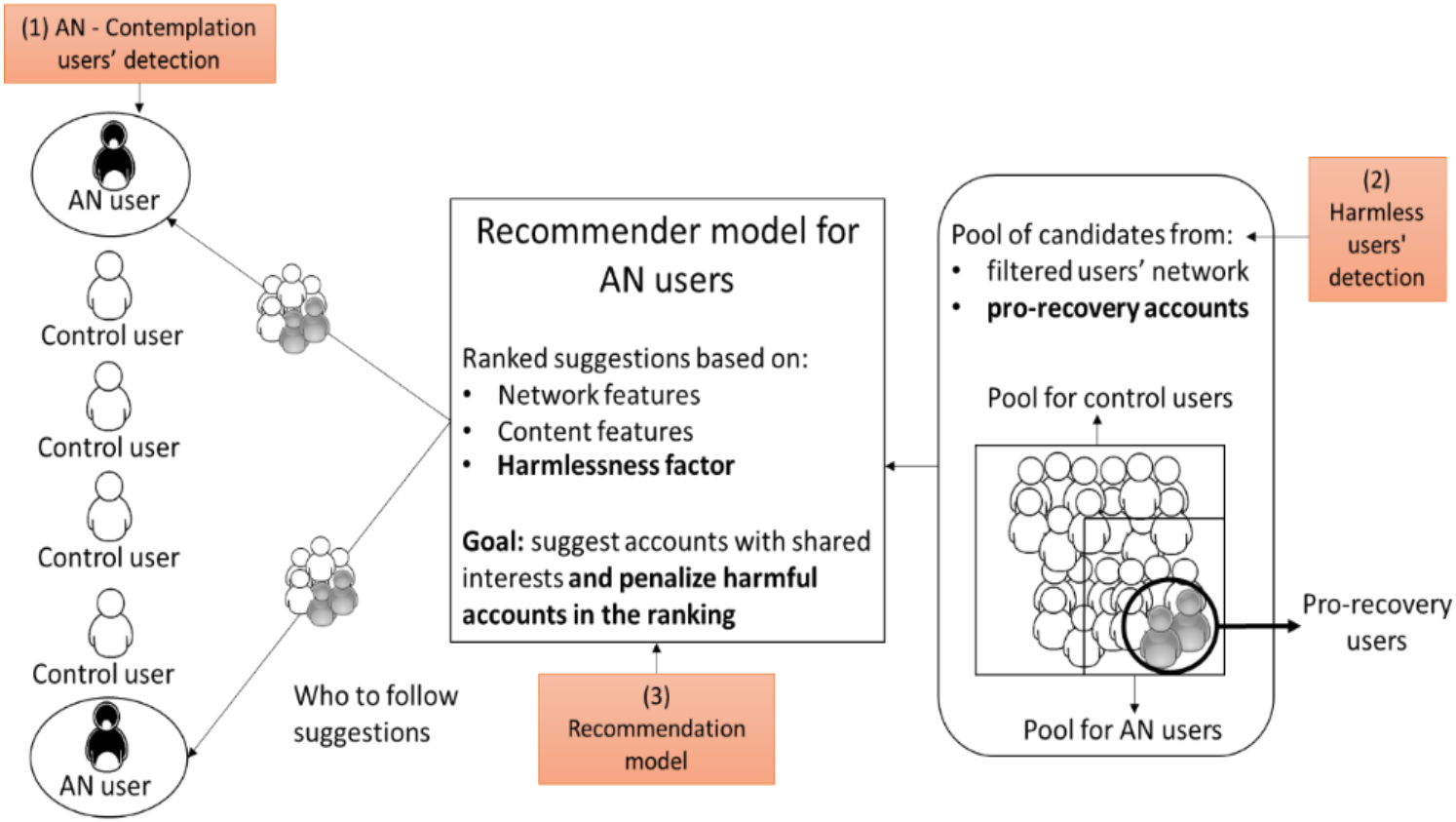

Figure 6.2. Architecture of the recommender system proposed.

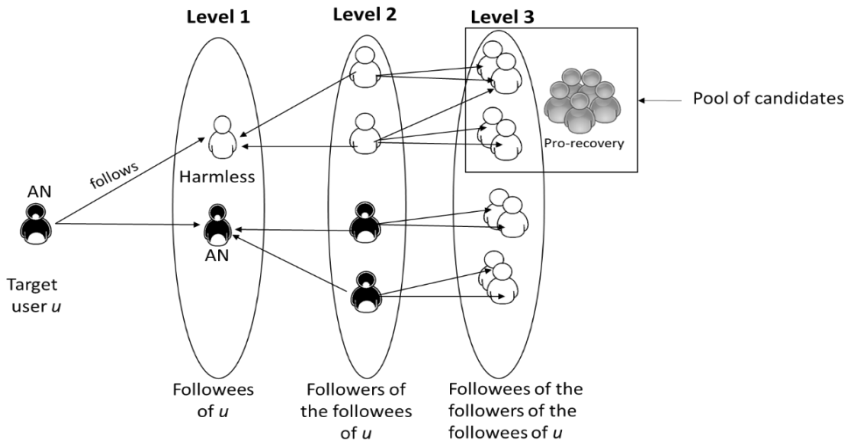

Figure 6.3. Definition of the pool of candidates for an AN target user u.

Considering that through this way most of the users in the pool would be harmful, we do a prior filtering step, where we apply a classifier to detect harmless users over u's followees (Level 1 users) so that the likelihood of suggesting harmful accounts is reduced.

Notice that through this study, asides from contemplation and control accounts, we also address 1) harmful accounts, which are those that can negatively influence the behavior of users with anorexia, here we can find accounts that promote diets and excessive exercising, accounts that express concerns about body image and promote unhealthy eating habits, and specially pro-ED accounts, among others. 2) Pro-recovery accounts that correspond to specialized recovery centers, educational psychologists, foundations and people that can offer support and information towards recovery from eating disorders. 3) Neutral accounts that do not promote harmful nor pro-recovery content. Finally, we consider 4) harmless accounts, which correspond to the union of neutral and pro-recovery accounts.

6.3 Detection of contemplation users

For the recommendation approach to target only users at the contemplation stage of AN, we developed a machine learning-based predictive model which was trained using features extracted from the sample of 56 users at the Contemplation

stage of AN (Dataset 3). This was assigned as the main target class to predict versus a group of control users.

Based in our predictive models described in Chapter 5, we evaluated multiple predictive models with several features including: 1) a TF-IDF [15] bag of words model (BoW), where users are represented by the frequency of the ({1-3}-grams) found in the writings. 2) A features' model named lexicon model where we extract attributes from the texts trying to map the characteristics that are often observed by clinicians for AN screening. The features were mainly gathered from the content shared and interests of the users. These features consist in linguistic and psychological aspects through the following categories: linguistic dimensions (24 features); affective processes and emotions (29 features); personal concerns and biological processes (12 features), vocabulary related to suicide risk factors (10 features) and vocabulary related to eating disorders (9 features).

Each of these models were tested using multiple classification methods such as Logistic regression (LR), random forest (RF), and Support vector machines (SVM), with 5-fold cross validation and applying oversampling methods to overcome imbalanced data issues.

The third predictive method 3) is a deep learning model that uses Convolutional Neural Networks (CNN) based on the approach described in [146]. This model uses word embeddings as the main input. The method was evaluated averaging the results of several runs, with a validation set of 10% of the training samples (70% of all the instances) in each run. All the models were evaluated in a test set which corresponds to the remaining 30% of all the data instances. We evaluated the performance of the models proposed in terms of Precision, Recall, and F1-Score for the main class to predict, and Accuracy for both classes.

The results for the prediction of the AN class are described in Table 6.1. The model selected as the best according to all measures is the BoW model with a LR classifier. This was therefore the model used for the detection of Contemplation users in the recommender evaluation. The performance of the BoW model suggests that the vocabulary used by contemplation users is quite distinguishable from the one of control users.

Table 6.1. Evaluation of contemplation users' detection model.

Model	Classifier	Precision	Recall	F1-Score	Accuracy
Bag of words model	Logistic Regression	0.94	**0.94**	**0.94**	**0.98**
Lexicon model	Random Forest	0.92	0.71	0.80	0.96
CNN model	CNN	**1**	0.40	0.57	0.94

6.4 Detection of harmless users

Since our approach filters user's followees to consider those that are more likely to be harmless for the pool of candidates, we created a classifier capable of distinguishing harmful from harmless accounts. To do so, using Dataset 3 – anorexia nervosa we labeled control accounts that included pro-recovery accounts among them (focused control accounts), as either harmless, harmful or doubtful (for those cases where annotators were not sure about their choice).

We also assigned automatically to the AN cases of the dataset (precontemplation and contemplation) the 'harmful' label. We then developed a harmful vs. harmless cases classifier.

We adopted the same approaches described for the contemplation users' detection to create our predictive models. The main target class assigned was the harmless one. The same evaluation approach and measures as for the contemplation users' classifier were used. This classifier is also used to calculate the harmlessness factor.

Our findings regarding the harmlessness classification model are described in Table 6.2. For this case, the lexicon model obtained the best results for all the evaluation measures, and thus became the model used for the recommendation approach.

The weakness of the BoW model may have been given by the fact that in the dataset there are harmless and harmful users that make use of AN vocabulary (i.e., AN and Pro-recovery users). Therefore, it is likely for the Lexicon model to have identified more attributes that characterize harmless from harmful accounts.

Table 6.2. Harmless users' detection models.

Model	Classifier	Precision	Recall	F1-Score	Accuracy
Bag of words model	SVM	0.69	0.85	0.76	0.78
Lexicon model	Random Forest	**0.79**	**0.95**	**0.86**	**0.87**
CNN model	CNN	0.68	0.74	0.71	0.74

6.5 Candidates ranking algorithm

Among the pool of candidates for a given target user u, we rank candidates based on a comparison between u and each of the candidates c_x to be recommended. We use similarity measures in order to suggest candidates that are more alike, in terms of shared interests (content), and the user's network topology [9,81,85]. In addition to these common elements, we add a harmlessness factor, which ranks recommendations based on how harmless for the user the candidate is likely to be. The elements considered to obtain a ranking score for each candidate, given a pair *(user u, candidate c_x)*, are defined by the following elements:

a) Topology attributes:

We take into account two elements: 1) as it is likely for users of level 2 (see Figure 6.3) to have followees in common, we measure the number of times the candidate c_x appears in the pool of candidates C_u of the user over the total number of existing candidates in C_u (Equation 6.1). Notice that for our experiments we defined each pool to have 100 random candidates among the eligible users. The next element is given by 2) the followees in common between c_x and u, which is defined by the calculation of Jaccard's similarity [15] between the set of followees of u and c_x (Equation 6.2). A similar method is used in [10] but they only consider the size of the intersection between the sets of followees of u and c_x. Finally, a topology score (Equation 6.3) is given by the average of both of these scores.

$$Ocurrences\_ratio(\mathrm{u}, c_x) = \frac{\#(c_x, C_u)}{|C_u|} \tag{6.1}$$

$$\mathrm{JSim}(\mathrm{u}, c_x) = \frac{|\mathrm{Followees}(\mathrm{u}) \cap \mathrm{Followees}(c_x)|}{|\mathrm{Followees}(\mathrm{u}) \cup \mathrm{Followees}(c_x)|} \quad (6.2)$$

$$Topology\_score(\mathrm{u}, c_x) = Avg(\mathrm{JSim}(\mathrm{u}, c_x), Ocurrences\_ratio(u, c_x)) \quad (6.3)$$

b) Content attributes:

We compare the interests of each candidate c_x with those of the target user u. Our goal is to recommend candidates that have more shared interests with u. Topics of interest are extracted using the approach described in Section 4.4.2 using Dandelion's entity API as the extraction tool. For the recommendation approach, the tool was chosen as an alternative to topic modeling approaches or TF-IDF representation of terms as it provides general yet specific enough categories that can represent the interests of a given user. Prior to the extraction of topics, we only consider writings with a positive polarity in order to keep only relevant terms and to avoid topics that may be mentioned but not liked by users. Later, we used a part of speech (POS) tagger to keep only nouns, verbs and adjectives as the terms from which we would extract topics. This topic extraction process is the one used through all our experimental setup.

Afterwards, topics of interest of all the users and candidates were represented in a bag of words (topics) model, where the score assigned to each (user, topic) was scaled between 0 and 1 based on the max and min scores for all the topics, according to each user.

It is important to mention that the model vocabulary was defined by the topics of interest of all the target users evaluated. With this, each user had a vector of topics representing their interest. This was done in order to compare the vector of topics of u denoted as v_u with the vector of topics of c_x denoted as v_{c_x} through the cosine similarity between them [9], as defined by Equation 6.4:

$$Content\_score\ (\mathrm{u}, c_x) = CosSim(v_u, v_{c_x}) = \frac{v_u . v_{c_x}}{||v_u|| \times ||v_{c_x}||} \quad (6.4)$$

c) Harmlessness factor:

We introduce a harmlessness factor, which penalizes harmful accounts in case they are part of the pool of candidates. This factor is given by a harmlessness score, which is represented by the output of the harmlessness classifier. The score is between [0,1] recalling that the higher the score, the less harmful the candidate is.

Finally, the rank score for u and c_x is given by Equation 6.5.

$$Rank\_score\ (u, c_x) = harmlessness\_score(c_x) \times Avg(Content\_score(\mathrm{u}, c_x), Topology\_score(\mathrm{u}, c_x)) \quad (6.5)$$

Notice that for the pro-recovery candidates, the rank score is given only by the product between the harmlessness and content scores.

6.6 Experimental framework

We evaluate the viability of our proposal with volunteers, further referred as survey participants, that have gone through the contemplation stage of AN. These are the same participants considered in the characterization of anorexia (Section 4.4.3). We also perform an annotation-based evaluation of the proposal, considering users' data.

6.6.1 Survey participants' evaluation

Recalling that volunteers participated without providing data from their social media accounts, the evaluation method consisted in 1) obtaining from participants (through surveys) a list of topics of their interest at the contemplation phase; 2) mapping the interests of participants to a proper format to compare them with the interests of Twitter candidates to recommend; 3) applying a variation of our rank score to recommend potential users to follow for each participant par (Equation 6.6); and 4) suggesting the top 5,10, and 15 candidates to participants and evaluating how likely they were to follow the users suggested through our approach.

$$Participants\_rank\_score(par, c_x) = harmlessness\_score(c_x) \times content\_score(par, c_x) \quad (6.6)$$

Notice that the pool of candidates for a target participant is given by the union of the pools of candidates of the Twitter users' evaluation approach (the methodology for defining these candidates is explained in Section 6.6.2). A set of 1,491 unique users were obtained.

Considering that we have previously obtained a vector of scored topics representing each target participant (process described in Section 4.4.3), we also proceeded to extract the candidates' topics from their Twitter profiles.

The frequency of appearance of each topic provided the vector of topics of the candidate. We applied the same normalization approach as for the participants.

Notice that we join all the topics from all participants for the further comparison between the vectors of topics of participants and candidates. We then applied the participants rank score (Equation 6.6) to rank candidates for each target participant.

a) Survey participants' evaluation baselines

In addition to our approach, we defined 5 baselines for recommending users with which we compare our recommendation approach. They are described in Table 6.3, where we show the recommendation methods, types of users of the pool of candidates, and ways for obtaining the pools of candidates.

We can see that the pool of candidates defined for model V.4 has several harmful candidates, while this changes when the filtering approach of our method is applied (model V.5).

Our model differs from model V.5 given that in addition to the content score, we consider the harmlessness score, precisely with the intention to rank at the top those harmless users that share interests with the participants. Moreover, our method introduces pro-recovery accounts in the pool of candidates given that it is less likely for these types of accounts to make it to the pool.

b) Survey participants' evaluation measures

To evaluate our model and baselines, we generate suggestions for participants, where for each model we show the top 15 candidates.

Table 6.3. Baselines defined for the evaluation of the participants.

Baseline model	Source of pool candidates	Types of users considered in the pool of candidates	Rank score per candidate of a given participant (par)
Model V.1	Sample of twitter accounts that were labeled as harmful.	Only harmful users	$Content\_score\ (\mathrm{par}, c_x)$
Model V.2	Sample of twitter accounts that were labeled as pro-recovery.	Only pro-recovery users	$Content\_score\ par$
Model V.3	Sample of twitter accounts that were labeled as either harmful, neutral or pro-recovery.	Equal number of pro-recovery, harmful and neutral users	Random suggestions
Model V.4	Sample of users obtained from the pool of candidates of the user's evaluation approach without considering the filtering step of our method. This would be equivalent to a state-of-the-art method of obtaining pool candidates.	Pro-recovery (0%), harmful (82%) and neutral (18%) users.	$Content\_score\ (\mathrm{par}, c_x)$
Model V.5	Sample of users obtained from the pool of candidates of the user's evaluation approach considering the filtering step of our method.	Pro-recovery (0%), harmful (30%) and neutral (70%) users.	$Content\_score\ (\mathrm{par}, c_x)$

Participants are expected to select the accounts they would have followed during the contemplation phase among those suggested. Notice that within the users suggested by our recommendation approach, 20% correspond to pro-recovery users as defined by the method proposed.

The evaluation measures are those commonly used for assessing recommender systems: precision (P), recall (R), and the mean average precision (MAP), all at K= 5, 10 and 15 recommendations [15]. Notice that for recall and precision we report the average of the results of all the participants. Also, given that we evaluated several models with the participants, they were only asked to choose who to follow among the top 15 users recommended by each model. In addition to these common measures that evaluate the likelihood of a participant to follow a recommended user, we also measure the ratios of harmful, neutral, pro-recovery and harmless (neutral + pro-recovery) users recommended by each model (#accounts of a given type suggested at K/K); along with the ratio of users of each of these types that would actually be followed over the number of suggested users of each type (#accounts of a given type followed at K/#accounts of a given type suggested at K). We also evaluate the ratio of accounts of each type followed at K (#accounts of a given type followed at K/K).

Finally, considering that a good recommendation model should maximize the average precision (AP) [15], and the ratio of harmless accounts followed for a given target user, we define an evaluation measure that aggregates both of these scores. The score denoted as the Average Precision-Harmlessness Ratio Score (APHR) for a target user is given by the harmonic mean between the average precision and the ratio of harmless users followed at K (# harmless users followed at K/K) denoted as HLFRK, as it can be seen in Equation 6.7. We consider the harmonic mean to be adequate as it would penalize strongly the cases where only harmful accounts are suggested. Also, to calculate this measure for all target participants or users, the MAP and the average of the HLFRK measure can be used instead.

$$APHR = 2 \times \frac{AP \times \mathrm{HLFRK}}{AP + \mathrm{HLFRK}} \tag{6.7}$$

c) Results

Results for the participants' evaluation are described in Table 6.4. We can observe the results for the baseline models defined, and our proposal.

We show results regarding Precision (P), Recall (R), Mean Average Precision (MAP), and pro-recovery suggested ratio (PRSR), neutral suggested ratio (NSR), harmful suggested ratio (HSR) and harmless suggested ratio (HLSR) of accounts at K.

We also report the ratio of followed pro-recovery (PRFRS), neutral (NFRS), harmful (HFRS) and harmless (HLFRS) accounts over the number of accounts suggested of each type at K. Finally, we calculate the ratio of followed pro-recovery (PRFRK), harmless (HLFRK) and harmful (HFRK) accounts over the total number of accounts suggested (k), as described in the Participants' recommendation evaluation measures section. We also calculate the Average Precision-Harmlessness Ratio Score (APHR).

Regarding Precision, it can be seen that baseline model V.5 performs better for every value of K, and also has the best MAP scores. However, this model does not take into account any pro-recovery candidates.

Regarding our approach, we can observe that there is a small difference in precision when compared with a model that only recommends harmful content (7% at worst, when K=5). However, our proposal outperforms model V.4, which is the most similar to a common recommendation approach. We achieve an improvement in precision of up to 3% and, moreover, our method does not suggest any harmful accounts.

Regarding recall (R), we can see that Models V.1, .5 and our proposed approach obtain the best results depending on the value of K.

Notice that, when K=15, recall is likely to be 1 as participants only annotated up to 15 suggestions per model. When it is not 1 is because R=0 when no relevant suggestions have been made.

Based on the ratio of pro recovery accounts suggested at K (PRSR), we can see that our model suggests most of the pro recovery accounts within the first 5 accounts suggested, and that 20% of these suggested accounts tend to be followed.

Table 6.4. Results for survey participants. We report Precision (P), Recall (R), Mean Average Precision (MAP), and pro-recovery suggested ratio (PRSR), neutral suggested ratio (NSR), harmful suggested ratio (HSR) and harmless suggested ratio (HLSR) of accounts at K accounts suggested. We also report the ratio of followed pro-recovery (PRFRS), neutral (NFRS), harmful (HFRS) and harmless (HLFRS) accounts over the number of accounts suggested of each type at K. We also calculate the ratio of followed pro-recovery (PRFRK), harmless (HLFRK) and harmful (HFRK) accounts over the total number of accounts suggested (k), along with the Average Precision-Harmlessness Ratio Score (APHR).

Model	Description	K	P	R	MAP	PRSR	NSR	HSR	HLSR	PRFRS	NFRS	HFRS	HLFRS	PRFRK	HLFRK	HFRK	APHR
Model V.1	Only harmful accounts + content	5	0.25	0.36	0.16	0.00	0.00	**1.00**	0.00	0.00	0.00	0.25	0.00	0.00	0.00	**0.25**	0.00
		10	0.25	**0.78**	0.27	0.00	0.00	**1.00**	0.00	0.00	0.00	**0.25**	0.00	0.00	0.00	0.25	0.00
		15	0.24	**1.00**	0.35	0.00	0.00	**1.00**	0.00	0.00	0.00	**0.24**	0.00	0.00	0.00	**0.24**	0.00
Model V.2	Only beneficial accounts + content	5	0.25	0.30	0.16	**1.00**	0.00	0.00	**1.00**	**0.25**	0.00	0.00	0.25	**0.25**	0.25	0.00	0.20
		10	0.25	0.66	0.29	**1.00**	0.00	0.00	**1.00**	**0.25**	0.00	0.00	0.25	**0.25**	**0.25**	0.00	0.27
		15	0.18	0.75	0.31	**1.00**	0.00	0.00	**1.00**	0.18	0.00	0.00	0.18	**0.18**	0.18	0.00	0.23
Model V.3	equal harmful, neutral and beneficial + random	5	0.08	0.22	0.14	0.38	0.47	0.15	0.85	0.04	0.08	0.00	0.09	0.03	0.08	0.00	0.10
		10	0.11	0.64	0.20	0.33	0.46	0.21	0.79	0.17	0.11	0.06	0.13	0.05	0.10	0.01	0.13
		15	0.13	**1.00**	0.26	**0.35**	0.40	0.25	0.75	0.14	0.10	0.20	0.12	0.05	0.09	0.03	0.13
Model V.4	no filtering step + content score	5	0.18	0.34	0.14	0.00	0.30	**0.70**	0.30	0.00	0.08	0.16	0.08	0.00	0.05	0.13	0.07
		10	0.18	0.58	0.21	0.00	0.24	**0.76**	0.24	0.00	0.08	0.19	0.08	0.00	0.03	**0.15**	0.05
		15	0.17	0.88	0.27	0.00	0.25	**0.75**	0.25	0.00	0.17	0.15	0.17	0.00	0.04	0.13	0.07
Model V.5	filtering step + content score	5	**0.38**	**0.39**	**0.29**	0.00	**0.80**	0.20	0.80	0.00	**0.26**	**0.33**	**0.26**	0.00	**0.23**	0.15	**0.26**
		10	**0.33**	0.67	**0.43**	0.00	**0.73**	0.28	0.73	0.00	**0.26**	**0.18**	**0.26**	0.00	0.21	0.11	**0.28**
		15	**0.27**	0.88	**0.48**	0.00	0.68	0.33	0.68	0.00	**0.22**	**0.19**	**0.22**	0.00	0.17	0.10	0.25
Model proposed	Filtering step + content score + harmlessness factor + beneficial accounts	5	0.18	**0.39**	0.23	**0.73**	0.28	0.00	**1.00**	**0.20**	0.06	0.00	0.18	0.15	0.18	0.00	0.20
		10	0.21	0.71	0.32	**0.36**	0.64	0.00	**1.00**	**0.20**	0.21	0.00	0.21	0.08	0.21	0.00	0.25
		15	0.20	**1.00**	0.38	0.24	**0.75**	0.00	**0.99**	**0.20**	0.20	0.00	0.20	0.05	**0.20**	0.00	**0.26**

Moreover, when only pro-recovery accounts are suggested (Model V.2) we can see that at K=10 25% (PRFRK) of the suggested accounts are likely to be followed, and among all the accounts shown (K=15) 18% (PRFRK) of the accounts were followed, meaning that users with AN are willing to follow pro-recovery accounts almost as much as they are willing to follow only harmful accounts at K=10 (25% for HFRK for Model 1), and the difference is only of a 6% at K=15.

About model V.4, which represents a common recommender system, we can see that it provides a high ratio of harmful accounts suggested (up to 75% at K=15) without obtaining better results in P, R or MAP compared to our proposal. Finally, considering the Average Precision-Harmlessness Ratio (APHR), model V.5 and the model proposed obtain the best results. We can also notice that Model V.1 obtains the worst results given that no harmless account is suggested by this model.

6.6.2 Twitter users' evaluation

We follow a similar approach as for the participants' evaluation but with contemplation Twitter users. The evaluation process consists in generating suggestions for a sample of contemplation target users through our recommendation approach. We start this evaluation by obtaining a new sample of users through the application of our contemplation users' detection approach.

As a first filtering approach for data collection, we obtain samples of user's timelines using keywords related to anorexia nervosa, the same keywords and phrases used for collecting Dataset 3 – anorexia nervosa.

A total of 773 profiles were collected. Over this sample of users, we applied the classifier dedicated to the detection of contemplation cases. Like this, we only kept those users for which the classifier obtained a predictive probability for the contemplation class over 0.95, which corresponded to a high level of certainty.

Among these users we kept the top 20 with the highest predictive probabilities, and in order to evaluate the recommendation approach, before using these user's posts for the evaluation, we verified that these 20 cases corresponded to contemplation cases.

The choice of using only 20 contemplation users for evaluation was the fact that for a given user, the pool of candidates to suggest is obtained from the social network of the user taking into account users from level 3 according to Figure 6.3. Therefore, for each user we collect a pool of 200 followees. Among these followees we apply a classifier that detects whether these followees are harmless.

Within those users that are harmless we choose 20 random users in order to gather 20 followers for each followee, and 20 followees per each follower, dealing with a total of 160,000 users.

Notice that from the 8,000 candidates per user, for the evaluation purposes we choose 100 random users to be part of the pool of candidates to be recommended to the user, as these users had to be manually labeled later for evaluation purposes.

The process followed for the users' evaluation was the following: 1) we extracted from the contemplation target users' posts a list of topics of interest 2) for each contemplation user we obtained its own pool of candidates, and extracted the topics of interest of the candidates. 2) Then we applied the harmlessness classifier over the candidates' data, and also obtained the list of followees of the candidates and contemplation users (to calculate Jaccard's similarity). We then 3) applied our rank score over each target contemplation user and each of its pool's candidates (Equation 6.5). Finally, 4) we suggested the top 5, 10, 15, and 20 candidates to each target user and evaluated how likely they are to follow the accounts suggested through our approach.

As the owners of the evaluation accounts (contemplation users) did not participate in the analysis, 3 annotators labeled the outputs of the recommendation models for each contemplation user, selecting those accounts that the user would be more likely to follow based on a description of their profile, and on the fact that they were at the contemplation stage.

We only marked as followed the candidates for which all the annotators agreed that the target user would follow. We selected this evaluation approach instead of other evaluation methods, which involve using a test set of followees of the user [68], given that it is not likely for users to be currently following pro-recovery accounts, and we already make use of the few harmless users followed to obtain the candidates.

a) Twitter Users' evaluation baselines

We compare the results of our approach with the baseline models described in Table 6.5. Notice that we consider Twitter's recommender system as another baseline, but only to compare the types of accounts suggested by the platform with the ones of our method.

b) Users' recommendation models evaluation measures

We use the same evaluation measures as for the survey participants' evaluation process. Notice that for model U.5 we only compare with our model the percentages of harmful, beneficial and neutral users suggested at K=50.

c) Results

Results addressing Twitter users' evaluation are described in Table 6.6. The highest precision and recall at any value for K are given by Model U.1 (common recommender). Despite these results, Model U.1 is also the one with the most harmful suggestions made and selected.

As for the survey participants' case, the model with the most neutral suggested users corresponds to model U.2, which applies the filtering step proposed by our approach.

The models with most pro-recovery users suggested are model U.4 (K=15, K=20) and the model proposed (K=5, K=10). However, as the recommendations provided by model U.4 are random, the ratio of accounts followed is lower than the ratio corresponding to the model proposed.

The model proposed obtains for all K values the highest number of harmless and pro-recovery followed accounts. Also, when comparing Model U.3, with the model proposed, there is an improvement in the general results provided by the usage of the network features, as it is the only difference among the models.

Finally, the APHR score favors the model proposed as it keeps a good balance between MAP and HLFRK. Also, the difference in precision between model U.1 and the model proposed (lowest value) at K=20 is 17%, which is acceptable considering the quality of the accounts suggested.

Table 6.5. Baselines defined to evaluate Twitter users' recommendation approach.

Baseline model	**Source of pool candidates for each user**	**Types of users considered in the pool of candidates**	**Rank score per candidate of a given user**
Model U.1	Sample of users obtained from the pool of candidates without considering the filtering step of our method.	Beneficial (0%), harmful (82%) and neutral (18%) users.	$Content\_score(\mathrm{u}, c_x)$
Model U.2	Sample of users obtained from the pool of candidates of the user's evaluation approach considering the filtering step of our method.	Beneficial (0%), harmful (30%) and neutral (70%) users.	$Content\_score(\mathrm{u}, c_x)$
Model U.3	Sample of users obtained from the pool of candidates of the user's evaluation approach considering the filtering step of our method.	Beneficial (0%), harmful (30%) and neutral (70%) users.	$harmlessness\_score(c_x) \times Content\_score(\mathrm{u}, c_x)$
Model U.4	Sample of twitter accounts that were labeled as either harmful, neutral or beneficial.	Equal number of beneficial, harmful and neutral users	Random suggestions
Model U.5	Twitter's recommender system's pool of candidates.	Unknown	Unknown

Table 6.6. Results obtained for the evaluation of the baselines and the proposed model for users.

Model	Description	K	P	R	MAP	PRSR	NSR	HSR	HLSR	PRFRS	NFRS	HFRS	HLFRS	PRFRK	HLFRK	HFRK	APHR
Model U.1	No filtering step + content score	5	**0.60**	**0.26**	0.21	0.00	0.08	**0.92**	0.08	0.00	0.13	**0.62**	0.13	0.00	0.03	**0.57**	0.05
		10	**0.61**	**0.52**	0.37	0.00	0.11	**0.90**	0.11	0.00	0.21	**0.65**	0.21	0.00	0.03	**0.58**	0.06
		15	**0.60**	**0.77**	0.54	0.00	0.12	**0.88**	0.12	0.00	0.15	**0.60**	0.15	0.00	0.02	**0.58**	0.04
		20	**0.58**	**1.00**	0.67	0.00	0.14	**0.86**	0.14	0.00	0.19	**0.63**	0.19	0.00	0.03	**0.55**	0.06
Model U.2	Filtering step + content score	5	0.50	0.27	0.19	0.07	**0.46**	0.47	0.53	0.03	**0.20**	0.44	0.21	0.01	0.18	0.32	0.18
		10	0.46	0.51	0.32	0.05	**0.49**	0.47	0.54	0.03	0.20	0.40	0.20	0.01	0.17	0.29	0.22
		15	0.46	0.75	0.44	0.04	**0.50**	0.46	0.54	0.03	0.21	0.42	0.21	0.00	0.16	0.30	0.23
		20	0.46	**1.00**	0.56	0.04	**0.51**	0.46	0.54	0.03	0.20	0.40	0.20	0.01	0.16	0.30	0.25
Model U.3	Filtering step + content score + harmlessness factor + beneficial accounts	5	0.34	0.25	0.16	0.76	0.17	0.07	0.93	0.31	0.11	0.20	0.32	0.25	0.30	0.04	0.21
		10	0.34	0.47	0.25	**0.45**	0.33	0.23	**0.78**	0.32	**0.23**	0.33	0.28	0.14	0.22	0.12	0.23
		15	0.36	0.74	0.35	0.30	0.43	0.27	0.73	0.32	0.28	0.34	0.31	0.09	0.22	0.14	0.27
		20	0.37	**1.00**	0.44	0.23	0.46	0.31	0.7	0.32	**0.28**	0.33	0.30	0.07	0.20	0.16	0.28
Model U.4	random recommendations	5	0.27	0.24	0.16	0.34	0.40	0.26	0.74	0.25	0.08	0.44	0.12	0.08	0.10	0.17	0.12
		10	0.26	0.49	0.24	0.33	0.42	0.26	0.75	0.24	0.09	0.53	0.14	0.07	0.11	0.15	0.15
		15	0.25	0.74	0.31	**0.31**	0.44	0.25	**0.75**	0.28	0.07	0.66	0.14	0.08	0.11	0.15	0.16
		20	0.25	0.95	0.37	**0.30**	0.45	0.25	**0.75**	0.29	0.07	0.61	0.14	0.08	0.11	0.14	0.17
Model proposed	Filtering step + content score + harmlessness factor + social network features + beneficial accounts	5	0.4	**0.26**	0.17	**0.81**	0.14	0.05	**0.95**	**0.44**	0.08	0.10	**0.40**	**0.36**	**0.38**	0.02	**0.23**
		10	0.36	0.46	0.26	**0.45**	0.34	0.22	**0.78**	**0.42**	0.19	0.27	**0.34**	**0.18**	**0.27**	0.10	**0.26**
		15	0.39	0.74	0.37	0.30	0.42	0.28	0.72	**0.42**	**0.31**	0.28	**0.37**	**0.12**	**0.26**	0.13	**0.31**
		20	0.41	**1.00**	0.48	0.23	0.46	0.31	0.69	**0.42**	**0.28**	0.33	**0.35**	**0.09**	**0.24**	0.17	**0.32**

For Twitter's recommendation approach at K=50, 73.70% of the users suggested are harmful, and only 1% are pro-recovery. Our proposal suggests less harmful accounts: 21% beneficial, 47% neutral and 32% harmful (K=50).

6.7 Discussion

In Chapter 6, we first found that users at the contemplation stage could be automatically detected through the definition of a classification approach based on a bag of words model, which obtained a F1 score of 0.94 for the detection of contemplation cases, which makes it a suitable method for such a task.

We also trained a model for the detection of harmless accounts. This model was based on several lexicon features, which obtained an 87% accuracy at distinguishing harmless from harmful accounts. This classifier offered promising results when used for the filtering approach, and for the objective function of the recommendation approach.

Notice that we could have skipped the filtering step so that only with the proposed objective function, harmless accounts would have still been placed on top. However, we find the filtering step to be relevant for efficiency purposes, as harmful users are likely to be following mostly harmful accounts.

We defined a recommendation method that minimizes the number of harmful users suggested in comparison to common recommendation approaches and to Twitter's recommendation service. It recommends 68% of harmless accounts in the worst case (k=50) at the Twitter users' evaluation, in comparison to a 25.30% of accounts suggested by Twitter (k=50), and a 14% of accounts suggested by the baseline model that imitates a common recommendation approach (k=20).

Complementing the previous results, we have found that for both cases (participants and users evaluation), people with AN at the contemplation stage are likely to follow harmless accounts, including pro-recovery users, proving that the implementation of such recommender systems is a valid approach to encourage people with eating disorders to seek for help.

Finally, we have defined the APHR measure, which seeks to evaluate a recommender system based on the average precision and the ratio of harmless accounts followed. This is relevant for designers in order to not only maximize the number of accounts followed by users, but also to give relevance to the selection of non-harmful accounts among the suggestions available.

Chapter 7

CONCLUSIONS AND FUTURE WORK

7.1 Introduction

Mental disorders are a serious health issue. Worldwide, until 2017, it was estimated that 792 million people lived with a mental health disorder, which corresponds to more than one in ten people globally (10.7%) [137]. A significant number of people with mental disorders receive no treatment for their condition given the limited access to mental health care facilities; the reduced availability of clinicians; the lack of awareness; and stigma, neglect, and discrimination surrounding mental disorders. In contrast, internet access and social media usage have increased significantly, providing experts and patients with a means of communication that may contribute to the development of methods for the detection of mental health issues in social media users.

In this chapter we summarize our contributions, analyze the limitations of our studies, and discuss open issues and future research directions on mental health assessment in social media.

7.2 Summary

Through this research work we have 1) provided insights about the behavior of users with mental disorders in social media (RQ1). 2) We have presented methods for the enhancement of text representations (word embeddings) adapted to binary and multiclass predictive tasks that address small data in specific

domains (RQ2). 3) We have developed several predictive models for the detection of mental disorders (RQ2); and 4) we have defined and evaluated a contact recommendation method dedicated to users with anorexia (RQ3). In this section we summarize the contributions of our research work (Chapters 4, 5 and 6) by answering our research questions (Chapter 3):

RQ1) which are the textual, visual, relational and behavioral elements that characterize disorders such depression, suicidal ideation, alcoholism and eating disorders like anorexia in social platforms?

We addressed this research question in Chapter 4. In Tables 7.1 and 7.2 we summarize our main findings regarding elements that characterize users with the mental health conditions studied.

We have first identified elements that distinguish mental disorders in general [127] from control cases in a social media context. Then we have explored elements that distinguish selected disorders (in particular anorexia [125] and suicidal ideation [126]) from control cases, taking into account 2 types of control groups: a focused control group that addresses topics related to the use case studied, and a random or generic control group, which is focused on topics that are not necessarily related to the use case. Finally, we have performed a comparative analysis among multiple conditions, seeking for elements capable of distinguishing one from another [127].

Among the disorders analyzed, we have delved into the study of anorexia performing an analysis of the stages towards recovery according to the Trans theoretical model of health behavior change [125]. Considering our findings regarding the polarization between AN and pro-recovery communities, and the shift in the interests of users as recovery progresses, we have also analyzed the influence of social recommendation engines on people with AN [124].

In our work, we analyzed content at a post and at a user level. We used data mainly from two different social platforms: Reddit and Twitter, with data collected in English and Spanish. We analyze language, behavioral, demographic, visual, and domain specific features.

Table 7.1 Summary of insights of the mental conditions characterized.

Health state	Dataset	Data source	Lang.	Main findings and contributions
Mental disorders	Dataset 4b - mental	Reddit	English	• The MEN group addresses mostly topics related to feelings and emotions compared to control cases, • Sadness and fear are the emotions that mostly characterize the MEN group compared to control cases. • There are significant differences for all the personal concerns and biological processes features. Features such as work, achievement, leisure, home, money and religion are mostly addressed by control cases. • Writings of users of the MEN group, in comparison to the CON group, tend to have more first-person singular pronouns, use more negations, adverbs, verbs, and past and present verb tenses. • Mental disorders, substance abuse related vocabulary and risk factors are highly referenced by the MEN group.
Anorexia	Dataset 3 - anorexia nervosa	Twitter	Spanish	• We address the stages towards recovery from AN using the TTM. • AN users have more activity at night and on weekends compared to control cases. • Visual features are relevant to distinguish AN users from random and focused control cases. • Recovered users make more use of first person plural pronouns compared to people at the early AN stages. • The proportion of tweets related to AN drops significantly as recovery progresses. • There is a strong polarization among AN and focused control (pro-recovery) communities. • There is a shift in the interests of users as recovery progresses, meaning that recovered users have more similar interests to focused control users than AN users. • In average, 73.70% of the accounts suggested by Twitter to AN users to follow are harmful. • A 77.27% of survey participants that are in treatment for AN think that the content suggested by social platforms is harmful for them. • The main topic of interest of people at the contemplation stage is related to AN (nutrition).
	Dataset 1b - anorexia	Reddit	English	• Terms related to signs and symptoms of anorexia (anorexia, anorexic, meal plan, underweight, eating disorders, diagnosed, macros, cal, etc.) are the terms that characterize AN cases
Depression	Dataset 4a - multiple	Reddit	English	• Characterized by references to antidepressants and depression itself. • It has the least features with significant differences when compared to alcoholism cases. • Anger, sadness and death related terms are expressed significantly more compared to the ED and ALC groups. • Compared to the ALC, ED and SUI groups, there is a higher usage of third person plural pronouns.

Table 7.2 Summary of insights of the mental conditions characterized.

Health state	Dataset	Data source	Lang.	Main findings and contributions
Depression	Dataset 4a - multiple	Reddit	English	• Compared to the suicidal ideation group, people with depression express more feelings of contentment, love and zest. They also address more topics related to daily activities such as white-collar job, occupation, and sports. • It is the group with the fewest predictive terms (depression, anxiety, depressed, energy, mental health, sad) characterizing the class compared to the SUI, ED and ALC groups.

Suicidal ideation	Dataset 2 - suicidal ideation	Twitter	Spanish	• Compared to control cases, the suicidal ideation groups has:
				  • a high usage of terms related to health and biological aspects compared to control cases.
				  • a high usage of singular first person personal pronouns and verbs conjugated in singular first person.
				  • shorter texts, less friends and more tweets posted at sleep time.
				• Images are relevant to distinguish cases at risk from generic control cases, but not from focused control cases.

	Dataset 4a - multiple	Reddit	English	• The SUI groups expresses the most negative emotions compared to the DEP, ED and ALC groups. • There is a high usage of terms related to death and sexuality by the SUI group. • It obtains the lowest scores in the work, achievement and leisure categories. • Characterized by the usage of first person singular pronouns. • Explicit usage of suicide related vocabulary. Among the top terms that characterized the SUI group compared to the DEP, ED, and ALC groups we find: kill, suicide, die, killing and live. • Characterized for addressing topics such as: death (kill), crime, prison, weapon, war, fight, aggression, negative emotions and hate.
Alcoholism	Dataset 4a - multiple	Reddit	English	• Usage of vocabulary related to biological processes, ingest, leisure and substance abuse. • Low usage of adverbs, and a high usage of articles and prepositions. • It has the lowest mean value for the hate category when compared to the SUI, DEP and ED groups. • The topics that characterize the ALC group are alcohol, liquid, party, and smell. • When the ED and ALC groups are compared, the ALC group addresses more leisure related topics. • Among the top predictive terms we find: *alcoholism, alcohol, alcoholic, drinking, drink* and *sober*.
Eating disorders	Dataset 4a - multiple	Reddit	English	• Usage of terms related to achievement, food and meals, caloric restriction, anorexia promotion, eat verb, body image, binge eating, body weight, compensatory behavior and laxatives. • The topics most addressed by the ED group are food, eating, cooking, restaurant, shopping and strength. • The most relevant terms that characterize the ED group compared to the other groups are eating, eating disorder, bulimia, purging, and ED.

We have found that mental conditions are mainly characterized by the usage of first person singular pronouns for both Spanish and English cases; and that there is vocabulary that characterizes each condition in particular.

We have also been able to trace in social media signs and symptoms that are relevant for clinicians during consultation. This has been done with the usage of lexicons (*e.g.,* risk factors), the analysis of posting patterns (*e.g.,* sleep time posting ratios), and the exploration of the social network of users (*e.g.,* communities detection and analysis of interest shared with followees).

Within our framework the characterization of mental disorders provide insights and relevant features that can serve as inputs for detection and risk aware recommendation tools to assist people with mental disorders. These are the aspects that we explore in RQ2 and RQ3.

RQ2) how can the features that characterize mental disorders be exploited for the development of new automated and explainable detection methods that can assist specialists to reach out to people at risk?

We generated predictive models that make use of the features analyzed for the characterization of mental disorders (Chapter 5). As described in Table 7.3, we have evaluated several predictive models that were trained using data from different sources and that have addressed binary and multiclass tasks.

We have also analyzed the contribution of multiple feature types, and we have taken into account early risk detection settings for anorexia and depression screening. Moreover, we have also explored gender biases in a predictive model dedicated to detect anorexia cases.

Exploring the most predictive features for the detection of suicidal ideation [126] and anorexia cases [149] we highlighted the importance of the interpretability of our features, considering elements that can be understood by clinicians and mapped to their screening practice.

The results of our experiments showed that within the types of features analyzed, there were multiple significant features that may lead to the detection of risk situations.

Table 7.3. Summary of the best predictive models evaluated for each task.

Task ID	Type of task	Classes addressed	Granularity level	Dataset	Data source	Best model proposed	Classifier	F1 Score
T1	binary	suicidal ideation vs. control (SPVC)	post level	dataset 2 - suicidal ideation	Twitter	bag of words model	LR	0.9
T2	binary	suicidal ideation vs. focused control	user level	dataset 2 - suicidal ideation	Twitter	Images + SNPSY model	SVM	0.86
T3	binary	suicidal ideation vs. Generic (random) control	user level	dataset 2 - suicidal ideation	Twitter	Images + SNPSY model	LR	0.88
T4	binary	anorexia vs. Control	post level	dataset 3 - anorexia nervosa	Twitter	word embeddings model	CNN	0.98
T5	binary	anorexia vs. Control	user level	dataset 1b - anorexia	Reddit	variation 2 (Predictive pairs embeddings (βP = 50, βN = 50) + GloVe embeddings)	MLP	0.78
T6	multiclass	DEP, SUI, ED and ALC	post level	dataset 4a - multiple	Reddit	embedding model 4 (GloVe's initial weights + predictive terms + retrofitting)	DL1 (CNN)	0.86
T7	binary	mental conditions vs. Control	post level	dataset 4b - mental	Reddit	embedding model 1 (GloVe's initial weights + predictive terms)	DL1 (CNN)	0.96
T8	binary (early risk detection)	depression vs. Control	user level	dataset 1a - depression	Reddit	model 1 (LIWC features)	LR	0.55
T9	binary (early risk detection)	anorexia vs. control	user level	dataset 1b - anorexia	Reddit	model 3 (LIWC + unigrams and bigrams + anorexia vocabulary + anorexia feature with weighted scores)	LR	0.73
T10	binary	contemplation vs. Control	user level	dataset 3 - anorexia nervosa	Twitter	bag of words model	LR	0.94
T11	binary	harmless vs. Harmful	user level	dataset 3 - anorexia nervosa	Twitter	lexicon model	RF	0.86

The most relevant features are based on the identification of textual and behavioral elements such as self-references, posting frequencies and in particular the usage of domain specific vocabulary that characterizes the class to predict.

Regarding the usage of images, we have found that their inclusion slightly increases the performance of predictive models for the detection of suicidal ideation when combined with textual cues.

The relevance of the domain specific vocabulary leads us to take advantage of this aspect and generate more suitable text representations (word embeddings) adapted to the predictive task addressed. To do so we take into account the terms (n-grams) that characterize a given class. In this sense, we proposed methods to generate enhanced word embeddings that outperformed the results obtained by word2vec and that are compatible with other embeddings generation and enhancement approaches like GloVe and Faruqui's [51] retrofitting proposal.

Exploring the best predictive models presented (Table 7.3), we can conclude that textual features either through open vocabulary representations (BoW or word embeddings), or lexicons are relevant for creating models with a good performance.

As we can see for T2, T3, T8 and T11, models that combine lexicon based features (*e.g.,* LIWC and domain specific vocabulary) and behavioral features (*e.g.,* posting frequencies and social network) perform better than open vocabulary models (baselines). The usage of these models can provide specialists explainable outputs when assisting them in screening tasks.

Regarding our framework, RQ2 is addressed by the detection module. Through the analysis of several predictive models, we could define models that can be used in early risk detection and recommender systems.

Through our experiments we have defined a model to detect cases of anorexia at the contemplation stage in order to define the target users for contact recommender system, addressed by RQ3.

RQ3) Can a social recommender system for users with anorexia nervosa connect them with pro-recovery communities so that users at risk are encouraged to seek help?

Yes. Based on our prior findings, users living with anorexia nervosa (AN) tend to seek accounts of peers that support their unhealthy habits. Contact recommendation systems can unintentionally reinforce such behaviors. To address this issue we have proposed and evaluated a contact recommendation approach dedicated to maximize the number of harmless users suggested.

Results show that AN users are willing to follow harmless accounts suggested in online platforms. There is a tradeoff in precision (P) when comparing the model proposed (P=0.41) with a regular recommendation approach (P=0.58). However, these results are promising considering that through the model proposed there is a 55% increase in the percentage of harmless accounts suggested. This is the first proposal that designs a social recommendation method dedicated to AN users, defining measures for its evaluation.

Our findings are relevant as they prove that contemplation users are likely to have a positive reaction towards recommendation approaches that reduce the exposure of vulnerable users to harmful content.

Considering our framework, this question addresses the module of contact recommendation. Going back to the data driven process this can be interpreted as the outcome of our findings in the prior modules. This tool is intended to assist users at risk by encouraging them to reach out for help.

7.3 Limitations

The analyses performed in this thesis present certain limitations that are mainly given by the structure of the social platforms analyzed, which do not provide explicit information regarding elements relevant to our analysis, such as the location, age, gender, or medical records of users. This is the main reason that has led us to infer information based on the analysis of the users' posts; therefore, the accuracy of our results is limited to the performance of the tools we have applied.

This applies to the demographic features inferred (age groups and gender); the analysis of elements that are related to the signs and symptoms of anorexia (terms related to risk factors and domain related vocabulary); image analysis tools; and the inference of aspects that involve the location of the user, such as the weekend tweet count ratio and the sleep period tweeting ratios. This last feature is calculated in a way that overcomes the issue of not knowing the difference in the posting time according to the user's time zone. This aspect is also an issue regarding the tweeting frequency in different periods of the year, as the seasons change according to the location of users.

We also considered the limitations owing to the accuracy of the translation of terms to English for the annotation and use of topic detection tools.

There are also limitations posed by the characteristics of users who have a preference for the studied platforms and that choose to make their tweets publicly available, which might differ from those that keep their profiles private.

It is important to recall that our study is limited to users who make use of Twitter and Reddit; therefore, the analysis of the behavior of users from other social platforms and even of people with mental disorders that do not have accounts on any social platforms is out of our reach.

Regarding the contact recommendation proposal, the main limitation was given by the difficulty of reaching people at the contemplation phase. At this stage (contemplation), people have not been diagnosed, and therefore organizations are not yet in contact with patients. Therefore, volunteers participating were people at the last stages of their treatment. Thus, the knowledge acquired during their treatment process might have influenced their survey answers. Regardless of this, we think that having involved patients that have gone through the contemplation phase has been relevant for the outcomes of the study. The same limitation was found for the users' evaluation, for which we have inferred through annotators the choices of users.

There might also be biases introduced by the annotators and survey participants based on their personal background and beliefs.

7.4 Impact and future work

In this section we assess the impact of the thesis as a multidisciplinary work. We also describe elements that can contribute to the further improvement of the framework.

This research work has implied an application of computer science to the health field. Thus it has an impact in the social, health and well-being areas (third United Nations' sustainable development goal); as well as in the computer science field. The findings of this research work can be useful to contribute to the understanding of how mental disorders are manifested in online social platforms.

We hope that in the future, a proper use of predictive and intervention tools is done so that they can contribute to the early detection and treatment of mental disorders. We expect the contributions of this work to be relevant for the promotion of campaigns related to the prevention of suicide and general awareness regarding the importance of mental health, in collaboration with specialized centers dedicated to these tasks. In particular, our contributions related to suicidal ideation have reached important media, especially in the Spanish-speaking community [47].

Addressing the impact in the computer science field, we have contributed with new datasets, and the application of methods and models to analyze the behavior, content and structure of groups of users with mental health issues in online social networks. We improve text representations adapted to the detection of mental disorders and develop predictive models that address different mental disorders and stages within them. We also define a contact recommendation approach that takes into account the harmfulness of recommendations for users with mental disorders. Some of the methods defined have the potential to be used for different domains and tasks.

Most of the outcomes of this work have been disseminated in top-tier journals and conferences centered on e-health, social computing, data mining, and information systems.

We believe that our findings are relevant to the development of predictive models that can assist specialists in the detection of users with mental health issues. These tools can filter risk cases in social platforms and display indicators of risk factors

as well as signs and symptoms that characterize mental disorders. Like this, experts (clinicians) can reach out more people at risk while having relevant information (indicators) concerning the risk status of a given user. This is a relevant aspect to study in the future as there has not been an evaluation of the usage of these models as assessment tools for experts [26].

Further work should also be done to address biases and fairness issues in predictive models dedicated to the detection of mental disorders. This is relevant prior to the deployment of such models, as they have to be reliable enough. Also, a careful risk-benefit assessment and a proper analysis of the applicable legal framework compliance should be done.

In addition, data from current popular social platforms should be studied. An instance is Instagram, which at the moment when our studies were performed had several restrictions regarding the access to their content through their API, but has recently released a new version of it. Another interesting platform is TikTok which became popular at the final stages of this research work. Even though it is not mainly used to express the current thoughts and moods of a user (our scope), it is relevant to study because of the diffusion it can give to harmful content. In this sense, the analysis of the content in videos represents a new challenge for our research area.

Finally, just in time adaptive interventions seem to be promising for going a step further from the main focus in detection models to actual interventions that make use of such models. Predictive models based on social media data can be a powerful resource to define further interventions depending on the levels of risk that users go through. Such predictive models can be embedded in apps that along with other types of data (GPS, heart rate, etc.) can provide further information to act in a better way.

REFERENCES

1. Abboute A, Boudjeriou Y, Entringer G, Azé J, Bringay S, Poncelet P. Mining Twitter for suicide prevention. In: Métais E, Roche M, Teisseire M, editors. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) Cham: Springer International Publishing; 2014. p. 250–253.
2. Agarwal S, Sureka A. Applying Social Media Intelligence for Predicting and Identifying On-line Radicalization and Civil Unrest Oriented Threats. 2015; abs/1511.06858.
3. Agència de Qualitat d'Internet (IQUA), Associació contra la Bulímia i l'Anorèxia (ACAB). Les pàgines "pro ana" i "pro mia" inunden la xarxa [Website]. 2011. Available from: http://www.f-ima.org/ca/que-fem/informes
4. AI Chooch. AI API | Computer Vision with Chooch AI [Website]. 2022 Available from: https://chooch.ai/api/#general-recognition-api
5. Al-Mosaiwi M, Johnstone T. In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation. Clinical Psychological Science 2018;6(4):529–542.
6. AlDayel A, Magdy W. Stance Detection on Social Media: State of the Art and Trends. Information Processing and Management; 2021; 58(4):102597.
7. Allen M. Intercoder Reliability Techniques: Percent Agreement. The SAGE Encyclopedia of Communication Research Methods; SAGE Publications, Inc; 2017.
8. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 5th ed. Washington: APA; 2013.
9. Armentano MG, Godoy D, Amandi A. Towards a Followee Recommender System for Information Seeking Users in Twitter. CEUR Workshop Proceedings Girona,

Spain; 2011.
10. Armentano MG, Godoy D, Amandi A. Topology-Based Recommendation of Users in Micro-Blogging Communities. Journal of Computer Science and Technology Springer; 2012; 27(3):624–634.
11. Arseniev-Koehler A, Lee H, McCormick T, Moreno MA. #Proana: Pro-Eating Disorder Socialization on Twitter. Journal of Adolescent Health; 2016; 58(6):659–664.
12. Attia E, Walsh BT. Anorexia Nervosa. American Journal of Psychiatry; American Psychiatric Association; 2007; 164(12):1805–1810.
13. Australian Government Department of Health. Department of Health | Suicidality. [Website] 2009. Available from: http://www.health.gov.au/internet/publications/publishing.nsf/Content/mental-pubs-m-mhaust2-toc~mental-pubs-m-mhaust2-hig~mental-pubs-m-mhaust2-hig-sui
14. Baeza-Yates R. BIG, small or Right Data: Which is the proper focus? [Website]. KGnuggets. 2018. Available from: https://www.kdnuggets.com/2018/10/big-small-right-data.html
15. Baeza-Yates RA, Ribeiro-Neto B. Modern Information Retrieval. 2nd ed. Harlow, England: Pearson Addison Wesley; 2011. ISBN: 0321416910, 9780321416919.
16. Bagroy S, Kumaraguru P, De Choudhury M. A Social Media Based Index of Mental Well-being in College Campuses. Conference on Human Factors in Computing Systems - Proceedings [Website] New York, NY, USA: ACM; 2017. p. 1634–1646.
17. Bastian M, Heymann S, Jacomy M. Gephi: An Open Source Software for Exploring and Manipulating Networks. Third International AAAI Conference on Weblogs and Social Media. 2009.
18. Baziotis C, Pelekis N, Doulkeridis C. DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) Vancouver, Canada; 2018. p. 747–754.
19. Benton A, Mitchell M, Hovy D. Multi-Task Learning for Mental Health using Social Media Text. Proceedings of

the 15th Conference of the EACL; 2017; abs/1712.0:152–162.
20. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast Unfolding of Communities in Large Networks. Journal of Statistical Mechanics: Theory and Experiment; 2008; 2008 (10).
21. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics 2017; 5:135–146.
22. Breiman L. Random forests. Machine Learning; 2001; 45(1):5–32.
23. Burger JD, Henderson J, Kim G, Zarrella G. Discriminating Gender on Twitter. EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference; Association for Computational Linguistics; 2011. p. 1301–1309.
24. Burnap P, Colombo G, Amery R, Hodorog A, Scourfield J. Multi-class Machine Classification of Suicide-related Communication on Twitter. Online Social Networks and Media; 2017; 2:32–44.
25. Chancellor S, Birnbaum ML, Caine ED, Silenzio VMB, De Choudhury M. A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media. FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency; USA: ACM; 2019. p. 79–88.
26. Chancellor S, De Choudhury M. Methods in Predictive Techniques for Mental Health Status on Social Media: a Critical Review. NPJ Digital Medicine 2020; 3(1).
27. Chancellor S, Mitra T, De Choudhury M. Recovery Amid Pro-anorexia: Analysis of Recovery in Social Media. Conference on Human Factors in Computing Systems - Proceedings 2016. p. 2111–2123.
28. Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 2002; 16:321–357.
29. Cheng Q, Li TM, Kwok CL, Zhu T, Yip PS. Assessing Suicide Risk and Emotional Distress in Chinese Social Media: A Text Mining and Machine Learning Study.

Journal of Medical Internet Research; 2017; 19(7).
30. Cheng Z, Caverlee J, Lee K. You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter users. International Conference on Information and Knowledge Management, Proceedings [Website] New York, NY, USA: ACM; 2010. p. 759–768.
31. Chitra U, Musco C. Understanding Filter Bubbles and Polarization in Social Networks. 2019; arXiv:1906.08772.
32. Cholbi M. Suicide: the Philosophical Dimensions. Choice Reviews Online. Broadview Press; 2012; ISBN: 1551119056.
33. De Choudhury M. Anorexia on Tumblr: A Characterization Study. ACM International Conference Proceeding Series; New York, NY, USA: ACM; 2015. p. 43–50.
34. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting Depression Via Social Media. Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013; AAAI; 2013. p. 128–137.
35. Cinelli M, de Francisci Morales G, Galeazzi A, Quattrociocchi W, Starnini M. The Echo Chamber Effect on Social Media. Proceedings of the National Academy of Sciences of the United States of America; National Academy of Sciences; 2021; 118(9).
36. Colombo GB, Burnap P, Hodorog A, Scourfield J. Analysing the Connectivity and Communication of Suicidal Users on Twitter. Computer Communications; 2016; 73:291–300.
37. Coppersmith DDL, Dempsey W, Kleiman EM, Bentley KH, Murphy SA, Nock MK. Just-in-Time Adaptive Interventions for Suicide Prevention: Promise, Challenges, and Future Directions. PsyArXiv; 2021.
38. Coppersmith G, Dredze M, Harman C, Hollingshead K, Mitchell M. CLPsych 2015 Shared Task: Depression and PTSD on Twitter. 2nd Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych 2015 - Proceedings of the Workshop Denver, Colorado; 2015. p. 31–39.
39. Coppersmith G, Leary R, Crutchley P, Fine A. Natural Language Processing of Social Media as Screening for Suicide Risk. Biomedical Informatics Insights; SAGE

Publications; 2018.
40. Cristianini, N; Ricci E. Support Vector Machines (SVM). In: Kao M-Y, editor. Boston, MA: Springer US; 2001. p. 349–361.
41. Davis R, Campbell R, Hildon Z, Hobbs L, Michie S. Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. Health Psychology Review; Taylor & Francis; 2015;9(3):323.
42. Deitrick W, Miller Z, Valyou B, Dickinson B, Munson T, Hu W. Gender Identification on Twitter Using the Modified Balanced Winnow. Communications and Network. 2012; 04(03):189–195.
43. Deriu J, Lucchi A, De Luca V, Severyn A, Muller S, Cieliebak M, Hofmann T, Jaggi M. Leveraging Large Amounts of Weakly Supervised Data for Multi-language Sentiment Classification. 26th International World Wide Web Conference, WWW 2017; 1045–1052.
44. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL HLT - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference; 2019; 1:4171–4186.
45. Divins M-J. Laxantes | Farmacia Profesional. Vol 30, Núm 4. p. 5–10.
46. Ebert DD, Harrer M, Apolinário-Hagen J, Baumeister H. Digital Interventions for Mental Disorders: Key Features, Efficacy, and Potential for Artificial Intelligence Applications. Advances in Experimental Medicine and Biology; Springer, Singapore; 2019;1192:583–627.
47. EFE A. Aplican la Inteligencia Artificial para Detectar Conductas Suicidas en la Red | Ciencia | Agencia EFE [Website]. 2021. Available from: https://www.efe.com/efe/espana/efefuturo/aplican-la-inteligencia-artificial-para-detectar-conductas-suicidas-en-red/50000905-4504737
48. Ellison NB, Boyd DM. Sociality Through Social Network Sites. The Oxford Handbook of Internet Studies. Oxford University Press; 2013. p. 151–172.

49. Elrod K, Dykeman C. A Corpus Linguistic Analysis of Pro-Anorexia Public Tumblr Posts Written in Spanish. Psychotherapy; 2019.
50. Farnadi G, De Cock M, Tang J, Moens MF. User Profiling through Deep Multimodal Fusion. WSDM 2018 - Proceedings of the 11th ACM International Conference on Web Search and Data Mining; New York, NY, USA: ACM; 2018. p. 171–179.
51. Faruqui M, Dodge J, Jauhar SK, Dyer C, Hovy E, Smith NA. Retrofitting word vectors to semantic lexicons. NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference ACL; 2015. p. 1606–1615.
52. Fast E, Chen B, Bernstein MS. Empath: Understanding Topic Signals in Large-scale Text. Conference on Human Factors in Computing Systems – Proceedings ACM; 2016; 4647–4657.
53. Fazelpour S, Danks D. Algorithmic Bias: Senses, Sources, Solutions. Philosophy Compass John Wiley & Sons, Ltd; 2021; 16(8):e12760.
54. Fichter MM, Quadflieg N. Mortality in Eating Disorders - Results of a Large Prospective Clinical Longitudinal Study. International Journal of Eating Disorders; 2016; 49(4):391–401.
55. Flach P. Performance Evaluation in Machine Learning: The Good, the Bad, the Ugly, and the Way Forward. Proceedings of the AAAI Conference on Artificial Intelligence Association for the Advancement of Artificial Intelligence (AAAI); 2019; 33(01):9808–9814.
56. Fleiss JL. Measuring Nominal Scale Agreement among many Raters. Psychological Bulletin; 1971; 76(5):378–382.
57. Fowler JC. Suicide Risk Assessment in Clinical Practice: Pragmatic Guidelines for Imperfect Assessments. Psychotherapy 2012; 49(1):81–90.
58. Golbeck J, Robles C, Turner K. Predicting Personality with Social Media. Conference on Human Factors in Computing Systems – Proceedings; New York, NY, USA: ACM; 2011. p. 253–262.

59. Goldberg SB, Lam SU, Simonsson O, Torous J, Sun S. Mobile Phone-based Interventions for Mental Health: A Systematic Meta-review of 14 Meta-analyses of Randomized Controlled Trials. PLOS Digital Health; Public Library of Science; 2022.
60. Grill G. Future Protest Made Risky: Examining Social Media Based Civil Unrest Prediction Research and Products. Computer Supported Cooperative Work: CSCW: An International Journal; Springer Science and Business Media B.V.; 2021; 30(5–6):811–839.
61. Grimaudo L, Song H, Baldi M, Mellia M, Munafò M. TUCAN: Twitter User Centric Analyzer. Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM; 2013. p. 1455–1457.
62. Guntuku SC, Preotiuc-Pietro D, Eichstaedt JC, Ungar LH. What Twitter Profile and Posted Images Reveal about Depression and Anxiety. Proceedings of the 13th International Conference on Web and Social Media, ICWSM. 2019.
63. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting Depression and Mental Illness on Social Media: an Integrative Review. Current Opinion in Behavioral Sciences; 2017; 18:43–49.
64. Gupta P, Goel A, Lin J, Sharma A, Wang D, Zadeh R. WTF: The Who to Follow Service at Twitter. WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web 2013; 505–514.
65. Gyarmati L, Trinh T. Measuring User Behavior in Online Social Networks. IEEE Network 2010; 24(5):26–31.
66. Hallgren KA. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutorials in Quantitative Methods for Psychology. Tutor Quant Methods Psychology; 2012; 8(1):23–34.
67. Han S. Googletrans, PyPI library [Website]. 2020; Available from: https://pypi.org/project/googletrans/
68. Hannon J, Bennett M, Smyth B. Recommending Twitter Users to Follow using Content and Collaborative Filtering Approaches. RecSys'10 - Proceedings of the 4th ACM Conference on Recommender Systems; New York, New York, USA: ACM Press; 2010; 199–206.

69. Hasler G, Delsignore A, Milos G, Buddeberg C, Schnyder U. Application of Prochaska's Transtheoretical Model of Change to Patients with Eating Disorders. Journal of Psychosomatic Research; Journal of Psychosomatic Research; 2004; 57(1):67–72.
70. Heckler WF, de Carvalho JV, Barbosa JLV. Machine Learning for Suicidal Ideation Identification: A Systematic Literature Review. Computers in Human Behavior; 2022; 128:107095.
71. Hilgart M, Thorndike FP, Pardo J, Ritterband LM. Ethical Issues of Web-based Interventions and Online Therapy. The Oxford Handbook of International Psychological Ethics; New York, NY, US: Oxford University Press; 2012. p.161–175.
72. Hoek HW. Incidence, Prevalence and Mortality of Anorexia Nervosa and Other Eating Disorders. Current Opinion in Psychiatry. 2006. p. 389–394.
73. Hofman E. Senti-py [Internet]. Available from: https://github.com/aylliote/senti-py
74. Hswen Y, Naslund JA, Brownstein JS, Hawkins JB. Monitoring Online Discussions about Suicide among Twitter Users with Schizophrenia: Exploratory Study. JMIR Mental Health; JMIR Publications Inc.; 2018 13;5(4):e11483.
75. Huang X, Zhang L, Chiu D, Liu T, Li X, Zhu T. Detecting Suicidal Ideation in Chinese Microblogs with Psychological Lexicons. Proceedings - 2014 IEEE International Conference on Ubiquitous Intelligence and Computing, 2014 IEEE International Conference on Autonomic and Trusted Computing, 2014 IEEE International Conference on Scalable Computing and Communications and Associated Symposia/Workshops, UIC-ATC-ScalCom; 2014. p. 844–849.
76. Hwang JD, Hollingshead K. Crazy Mad Nutters: The Language of Mental Health. Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych 2016 at the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2016 San Diego, California; 2016. p. 52–62.

77. Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. PLoS ONE Public Library of Science; 2014; 9(6):e98679.
78. Jett S, La Porte DJ, Wanchisn J. Impact of Exposure to Pro-eating Disorder Websites on Eating Behaviour in College Women. European Eating Disorders Review European Eating Disorders Review; 2010; 18(5):410–416.
79. Jin L, Chen Y, Wang T, Hui P, Vasilakos A V. Understanding User Behavior in Online Social Networks: A survey. IEEE Communications Magazine 2013; 51(9):144–150.
80. Kang K, Yoon C, Kim EY. Identifying Depressive Users in Twitter using Multimodal Analysis. 2016 International Conference on Big Data and Smart Computing, BigComp 2016 Institute of Electrical and Electronics Engineers Inc.; 2016. p. 231–238.
81. Kaur K, Dhindsa KS. Classification of Followee Recommendation Techniques in Twitter. In: Bi Y, Bhatia R, Kapoor S, editors. Advances in Intelligent Systems and Computing Cham: Springer International Publishing; 2020. p. 527–540.
82. Kim Y. Convolutional Neural Networks for Sentence Classification. EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference. 2014.
83. Kruskal WH, Wallis WA. Use of Ranks in One-Criterion Variance Analysis. Journal of the American Statistical Association JSTOR; 1952; 47(260):583.
84. Kumar V, L. M. Predictive Analytics: A Review of Trends and Techniques. International Journal of Computer Applications Foundation of Computer Science; 2018; 182(1):31–37.
85. Kywe SM, Lim EP, Zhu F. A Survey of Recommender Systems in Twitter. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) Springer; 2012; 7710:420–433.
86. Lederman R, Wadley G, Gleeson J, Bendall S, Alvarez-

Jimenez M. Moderated Online Social Therapy: Designing and Evaluating Technology for Mental Health. ACM Transactions on Computer-Human Interaction; 2014; 21(1).
87. Leis A, Ronzano F, Mayer MA, Furlong LI, Sanz F. Detecting Signs of Depression in Tweets in Spanish: Behavioral and Linguistic Analysis. Journal of Medical Internet Research; 2019; 21(6).
88. Levi G, Hassner T. Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns. ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction; 2015]; p.503–510.
89. Lin H, Jia J, Qiu J, Zhang Y, Shen G, Xie L, Tang J, Feng L, Chua TS. Detecting Stress Based on Social Interactions in Social Networks. IEEE Transactions on Knowledge and Data Engineering IEEE Computer Society; 2017; 29(9):1820–1833.
90. López-Úbeda P, Plaza-Del-Arco FM, Díaz-Galiano MC, Alfonso Ureña-López L, Martín-Valdivia MT. Detecting Anorexia in Spanish Tweets. International Conference Recent Advances in Natural Language Processing, RANLP; 2019; p. 655–663.
91. Losada DE, Crestani F. A Test Collection for Research on Depression and Language Use. In: Fuhr N, Quaresma P, Gonçalves T, Larsen B, Balog K, Macdonald C, Cappellato L, Ferro N, editors. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) Cham: Springer International Publishing; 2016. p. 28–39.
92. Losada DE, Crestani F, Parapar J. CLEF 2017 eRisk overview: Early Risk Prediction on the Internet: Experimental Foundations. In: Linda Cappellato Nicola Ferro LGTM, editor. CEUR Workshop Proceedings CEUR-WS.org; 2017.
93. Losada DE, Crestani F, Parapar J. Overview of eRisk -- Early Risk Prediction on the Internet. Experimental IR Meets Multilinguality, Multimodality, and Interaction Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018) Avignon, France; 2018.
94. Mowafy M, Rezk A, El-bakry HM. An Efficient

Classification Model for Unstructured Text Document. American Journal of Computer Science and Information Technology 2018; 06(01).
95. Maia M, Almeida J, Almeida V. Identifying User Behavior in Online Social Networks. Proceedings of the 1st Workshop on Social Network Systems, SocialNets'08 - Affiliated with EuroSys 2008; New York, NY, USA: ACM; 2008. p. 13–18.
96. Mann HB, Whitney DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics. Institute of Mathematical Statistics; 1947; 18(1):50–60.
97. Marius-Constantin P, Balas VE, Perescu-Popescu L, Mastorakis N. Multilayer Perceptron and Neural Networks. WSEAS Transactions on Circuits and Systems Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS); 2009; 8(7):579–588.
98. Masood R, Ramiandrisoa F, Aker A. UDE at Erisk 2019: Early Risk Prediction on the Internet. CEUR Workshop Proceedings; 2019.
99. Masuda N, Kurahashi I, Onari H. Suicide Ideation of Individuals in Online Social Networks. PLoS ONE Public Library of Science; 2013; 8(4):e62262.
100. McHugh ML. Interrater reliability: the kappa statistic. Biochemia Medica; Croatian Society for Medical Biochemistry and Laboratory Medicine; 2012; 22(3):276.
101. McKenna SP, Doward LC, Davey KM. The Development and Psychometric Properties of the MSQOL. Clinical Drug Investigation. 1998. p. 413–423.
102. Mewton L, Andrews G. Cognitive Behaviour Therapy via the Internet for Depression: A Useful Strategy to Reduce Suicidal Ideation. Journal of Affective Disorders; Journal of Affective Disorders; 2015; 170:78–84.
103. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings; 2013; abs/1301.3.
104. Mohammad SM, Turney PD. Crowdsourcing a Word-emotion Association Lexicon. Computational Intelligence;

2013.
105. Moore TJ, Mattison DR. Adult Utilization of Psychiatric Drugs and Differences by Sex, Age, and Race. JAMA Internal Medicine; American Medical Association; 2017; 177(2):274–275.
106. Mrazek P. J. and Haggerty R. J. Risk and Protective Factors for the Onset of Mental Disorders. Reducing Risks for Mental Disorders: Frontiers for Preventive Research; NCBI Bookshelf; National Academies Press (US); 1994; (Dc):1–60.
107. Murphy CA. The Role of Perception in Age Estimation. Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering; Springer, Berlin, Heidelberg; 2012. p. 1–16.
108. O'Dea B, Wan S, Batterham PJ, Calear AL, Paris C, Christensen H. Detecting Suicidality on Twitter. Internet Interventions; Elsevier; 2015; 2(2):183–188.
109. Padez-Vieira F, Afonso P. Sleep Disturbances in Anorexia Nervosa. Advances in Eating Disorders; Informa UK Limited; 2016; 4(2):176–188.
110. Paffard M. Suicidal Ideation. Acute Medicine: A Symptom-Based Approach; 2014. p. 415–420.
111. Pantic I. Online Social Networking and Mental Health. Cyberpsychology, Behavior and Social Networking 2014; 17(10):652–657.
112. Parapar J, Martín-Rodilla P, Losada DE, Crestani F. eRisk 2021: Pathological Gambling, Self-harm and Depression Challenges. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2021; 12657 LNCS: 650–656.
113. Park M, Cha C, Cha M. Depressive Moods of Users Portrayed in Twitter. Proceedings of the ACM SIGKDD Workshop; on Healthcare Informatics Beijing, China; 2012. p. 1–8.
114. Pastore M, Calcagnì A. Measuring Distribution Similarities between Samples: A Distribution-free Overlapping Index. Frontiers in Psychology; Frontiers in Psychology; 2019; 10.
115. Pedregosa, F; Varoquaux, G; Gramfort, A; Michel, V; Thirion, B; and Grisel, O; and Blondel; and Prettenhofer

P, and Weiss, R; and Dubourg, V; and Vanderplas, J; and Passos A; and Cournapeau, D; and Brucher, M. and Perrot, M; and Duchesnay E; Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research; 2011; 12:2825–2830.
116. Pennington J, Socher R, Manning CD. GloVe: Global Vectors for Word Representation. EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference Association for Computational Linguistics; 2014. p. 1532–1543.
117. Pérez NP, Guevara López MA, Silva A, Ramos I. Improving the Mann-Whitney Statistical Test for Feature Selection: An Approach in Breast Cancer Diagnosis on Mammography. Artificial Intelligence in Medicine; Elsevier; 2015; 63(1):19–31.
118. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep Contextualized Word Representations. NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 2018. p. 2227–2237.
119. Plutchik R. A Psychoevolutionary Theory of Emotions. Social Science Information; SAGE Publications Inc.; 1982; 21(4–5):529–553.
120. Preoţiuc-Pietro D, Eichstaedt J, Park G, Sap M, Smith L, Tobolsky V, Schwartz HA, Ungar L. The Role of Personality, Age and Gender in Tweeting about Mental Illnesses. 2nd Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych 2015 - Proceedings of the Workshop Association for Computational Linguistics; 2015. p. 21–30.
121. Prieto VM, Matos S, Álvarez M, Cacheda F, Oliveira JL. Twitter: A Good Place to Detect Health Conditions. PLoS ONE; 2014; 9(1).
122. Prochaska JO, Velicer WF. The Transtheoretical Model of Health Behavior Change. American Journal of Health Promotion; 1997; 12(1):38–48.
123. PyPI. Pillow; PyPI. [Website]; 2021; available from: https://pypi.org/project/Pillow/

124. Ramírez-Cifuentes Diana; Freire Ana; Baeza-Yates Ricardo. A Contact Recommender System for Users with Anorexia Nervosa. Submitted; 2022.
125. Ramírez-Cifuentes D, Freire A, Baeza-Yates R, Lamora NS, Álvarez A, González-Rodríguez A, Rochel ML, Vives RL, Velazquez DA, Gonfaus JM, Gonzàlez J. Characterization of Anorexia Nervosa on Social Media: Textual, Visual, Relational, Behavioral, and Demographical Analysis. Journal of Medical Internet Research; 2021; 23(7).
126. Ramírez-Cifuentes D, Freire A, Baeza-Yates R, Puntí J, Medina-Bravo P, Velazquez DA, Gonfaus JM, Gonzàlez J. Detection of Suicidal Ideation on Social Media: Multimodal, Relational, and Behavioral Analysis. Journal of Medical Internet Research; 2020; 22(7):e17758.
127. Ramírez-Cifuentes D, Largeron C, Tissier J, Baeza-Yates R, Freire A. Enhanced Word Embedding Variations for the Detection of Substance Abuse and Mental Health Issues on Social Media Writings. IEEE Access; 2021; 9:130449–130471.
128. Ramírez-Cifuentes D, Largeron C, Tissier J, Freire A, Baeza-Yates R. Enhanced Word Embeddings for Anorexia Nervosa Detection on Social Media. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer; 2020; p. 404–417.
129. Ramírez-Cifuentes D, Mayans M, Freire A. Early Risk Detection of Anorexia on Social Media. In: Bodrunova SS, editor. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) Cham; 2018. p. 3–14.
130. Ramírez-Esparza N, Pennebaker JW, García FA, Suriá R. La psicología del Uso de las Palabras: Un Programa de Computadora que Analiza Textos en Español. Revista Mexicana de Psicología Sociedad Mexicana de Psicología; 2007; 24(1):85–99.
131. Reece AG, Danforth CM. Instagram Photos Reveal Predictive Markers of Depression. EPJ Data Science; 2017; 6(1):15.
132. Reece AG, Reagan AJ, Lix KLM, Dodds PS, Danforth CM, Langer EJ. Forecasting the Onset and Course of

Mental Illness with Twitter Data. Scientific Reports; Nature Publishing Group; 2017; 7(1):13006.
133. Ridout B, Campbell A. The Use of Social Networking Sites in Mental Health Interventions for Young People: Systematic Review. Journal of Medical Internet Research; JMIR Publications Inc.; 2018; 20(12).
134. Ríssola EA, Losada DE, Crestani F. A Survey of Computational Methods for Online Mental State Assessment on Social Media. ACM Transactions on Computing for Healthcare; ACM; New York, NY, USA; 2021; 2(2):1–31.
135. Ríssola EA, Aliannejadi M, Crestani F. Beyond Modelling: Understanding Mental Disorders in Online Social Media. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer; 2020. p. 296–310.
136. Ríssola E, Ramírez-Cifuentes D, Freire A, Crestani F. Suicide Risk Assessment on Social Media: USI-UPF at the CLPsych 2019 Shared Task. In Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology. Association for Computational Linguistics; 2019; p.167–171
137. Ritchie H, Roser M. Mental Health - Our World in Data. Mental Health. 2020 [Website]. Available from: https://ourworldindata.org/mental-health#citation
138. Rodriguez P, Gonzàlez J, M. Gonfaus J, Xavier Roca F. Integrating Vision and Language in Social Networks for Identifying Visual Patterns of Personality Traits. International Journal of Social Science and Humanity; 2019; 9(1):6–12
139. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR 2019;abs/1910.0. Available from: http://arxiv.org/abs/1910.01108
140. Sanz-Cruzado J, Castells P. Information Retrieval Models for Contact Recommendation in Social Networks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer, Cham; 2019; 11437 LNCS: 148–163.
141. Savand A. Stop-words; Python library [Website].

Available from: https://pypi.org/project/stop-words/
142. Sayers J. Mental Health: New Understanding, New Hope. WHO, editor. Bulletin of the World Health Organization; 2001; 79, 11:21.
143. Schutt R, O'Neil C. Doing Data Science: Straight Talk. 2013. ISBN: 9781449358655.
144. Schwartz HA, Eichstaedt J, Kern ML, Park G, Sap M, Stillwell D, Kosinski M, Ungar L. Towards Assessing Changes in Degree of Depression through Facebook. Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality Baltimore, Maryland USA; 2015. p. 118–125.
145. Shen G, Jia J, Nie L, Feng F, Zhang C, Hu T, Chua TS, Zhu W. Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution. IJCAI International Joint Conference on Artificial Intelligence International Joint Conferences on Artificial Intelligence; 2017; 0:3838–3844.
146. Shing HC, Nair S, Zirikly A, Friedenberg M, Daumé H, Resnik P. Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings. Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, CLPsych 2018 at the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HTL; 2018.
147. Shrout PE, Fleiss JL. Intraclass Correlations: Uses in Assessing Rater Reliability. Psychological Bulletin; 1979; 86(2):420–428.
148. Silvester S, Tanbakuchi A, Müller P, Nunez-Iglesias J, Harfouche M, Klein A, McCormick M, OrganicIrradiation, Rai A, Ladegaard A, Lee A, Smith TD, Vaillant GA, jackwalker64, Nises J, rreilink, Kemenade H van, Dusold C, Kohlgrüber F, Yang G, Inggs G, Singleton J, Schambach M, Hirsch M, Komarčević M, Niklas Rosenstein, Hsieh P-C, Zulko, Barnes C, Elliott A. imageio/imageio v0.9.0. 2020.
149. Solans D; Ramírez-Cifuentes D; Ríssola E; Freire A. Gender Bias when using Artificial Intelligence to assess Anorexia Nervosa on Social Media. Submitted; 2022.

150. Squicciarini A, Griffin C. Why and how to Deceive: Game Results with Sociological Evidence. Social Network Analysis and Mining; 2014; 4(1):1–13.
151. Statsmodels - proportions_ztest — [Website]. 2022. Available from: https://www.statsmodels.org/stable/generated/statsmodels.stats.proportion.proportions_ztest.html
152. Su C, Xu Z, Pathak J, Wang F. Deep Learning in Mental Health Outcome Research: a Scoping Review. Translational Psychiatry; Nature Publishing Group; 2020; 10(1):1–26.
153. Tamburrini N, Cinnirella M, Jansen VAA, Bryden J. Twitter Users Change Word Usage According to Conversation-partner Social Identity. Social Networks; 2015; 40:84–89.
154. Tausczik YR, Pennebaker JW. The Psychological Meaning of Words: LIWC and Computerized text analysis methods. Journal of Language and Social Psychology; 2010; 29(1):24–54.
155. Thompson S. Suicide and the Internet. Psychiatric Bulletin 2001; 25(10):400.
156. Tissier J, Gravier C, Habrard A. Dict2vec : Learning Word Embeddings using Lexical Dictionaries. EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings. 2017.
157. Tsugawa S, Kikuchi Y, Kishino F, Nakajima K, Itoh Y, Ohsaki H. Recognizing Depression from Twitter Activity. Conference on Human Factors in Computing Systems – Proceedings. New York, NY, USA: ACM; 2015. p. 3187–3196.
158. Tuna T, Akbas E, Aksoy A, Canbaz MA, Karabiyik U, Gonen B, Aygun R. User Characterization for Online Social Networks. Social Network Analysis and Minining; Springer Vienna; 2016; 6(1):104.
159. Twitter Development Platform. Sample Realtime Tweets [Website]. Twitter, editor. 2020. Available from: https://developer.twitter.com/en/docs/tweets/sample-realtime/api-reference/get-statuses-sample
160. U.S. Department of Health and Human Services. Designated Health Professional Shortage Areas Statistics Third Quarter of Fiscal Year 2020 Designated

HPSA Quarterly Summary As of June 30 2020; Bureau of Health Workforce Health Resources and Services Administration (HRSA) Designated Health Profession; 2020; 1–15.
161. Vedula N, Parthasarathy S. Emotional and Linguistic Cues of Depression from Social Media. ACM International Conference Proceeding Series Association for Computing Machinery; 2017; Part F128634:127–136.
162. Walsh CG, Chaudhry B, Dua P, Goodman KW, Kaplan B, Kavuluru R, Solomonides A, Subbian V. Stigma, Biomarkers, and Algorithmic Bias: Recommendations for Precision Behavioral Health with Artificial Intelligence. JAMIA Open; Oxford Academic; 2021; 3(1):9–15.
163. Wang T, Brede M, Ianni A, Mentzakis E. Social Interactions in Online Eating Disorder Communities: A Network Perspective. PLoS ONE Public Library of Science; 2018.
164. Wang T, Brede M, Ianni A, Mentzakis E. Characterizing Dynamic Communication in Online Eating Disorder Communities: A Multiplex Network Approach. Applied Network Science; Springer Science and Business Media LLC; 2019; 4(1):1–22.
165. Wang Y, Tang J, Li J, Li B, Wan Y, Mellina C, O'Hare N, Chang Y. Understanding and Discovering Deliberate Self-harm Content in Social Media. 26th International World Wide Web Conference, WWW; 2017; 93–102.
166. Wang Z, Hale SA, Adelani D, Grabowicz PA, Hartmann T, Flöck F, Jurgens D. Demographic Inference and Representative Population Estimates from Multilingual Social Media Data. The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019; New York, New York, USA: Association for Computing Machinery, Inc; 2019. p. 2056–2067.
167. World Health Organization (WHO). Suicide rate estimates, age-standardized - Estimates by WHO region. World Health Organization; 2019. [Website] Available from: http://apps.who.int/gho/data/view.main.MHSUICIDEREGv?lang=en
168. Withers SD. Categorical Data Analysis. International Encyclopedia of Human Geography; Wiley; 2009. p. 456–

462.
169. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional Neural Networks: an Overview and Application in Radiology. Insights into Imaging; 2018; 9(4): 611–629.
170. Yin W, Schütze H. Learning Word Meta-embeddings. 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers Berlin, Germany: Association for Computational Linguistics; 2016. p. 1351–1360.
171. Zheng X, Han J, Sun A. A Survey of Location Prediction on Twitter. IEEE Transactions on Knowledge and Data Engineering IEEE Computer Society; 2018; 30(9):1652–1671.
172. Zheng Y, Li L, Zhang J, Xie Q, Zhong L. Using Sentiment Representation Learning to Enhance Gender Classification for User Profiling. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); 2019; 11642 LNCS:3–11.
173. Zuluaga A. Ethics and suicidal behaviors. Colombian Psychiatry Journal 2001; 30(4).