Probabilistic Models and Natural Language Processing in Health

by

Aurora Cobo Aguilera

A dissertation submitted by in partial fulfillment of the requirements for the degree of Doctor of Philosophy in

Multimedia and Communications

Universidad Carlos III de Madrid

Advisors:

Antonio Artés Rodríguez
Pablo Martínez Olmos

December 2022

A ellos, siempre a ellos, mamá y papá.

ACKNOWLEDGEMENTS

Parece paradójico que haya acabado de escribir esta tesis cuya aplicación principal es la psiquiatría. Y lo digo porque justo empecé a maquetarla hace cuatro años, en el que considero que ha sido el momento más complicado de mi vida. La salud mental es un bien que poco se valora cuando se tiene y seguramente sea uno de los pilares fundamentales de la vida para desarrollarse como persona y estar en armonía. Por ese motivo, la aplicación que presento, la hago con un especial cariño. Y por ese motivo también, el día que decidí darme la oportunidad de arrancar y sacarla adelante, también me propuse un objetivo primordial. No quería hacer la mejor tesis del mundo, ni alcanzar la conferencia más importante del momento. Simplemente, me propuse hacerla con pasión y sobre todo disfrutarla. Era lo que me pedía la vida. Me lo tomé como un largo aprendizaje. Quise tocar muchas áreas dentro del *machine learning* porque sólo pensaba en aprender. Aprender conceptos y herramientas nuevas, enseñárselos a los demás a través de la docencia y aportar un granito de arena al mundo de la ciencia.

Tras esta reflexión, sinceramente siento, que AHORA es el momento de apuntar más alto, de seguir creciendo en todos los sentidos y de demostrar lo que tengo dentro y lo que he aprendido. Cuando eliges el camino del doctorado, sientes una pasión innata a esto y para mí el camino del aprendizaje y la investigación no acabará nunca.

Seguramente ésta sea una sección de agradeciemientos poco común, pero era la parte a la que más ganas le tenía para escribir. Yo la consideraría la sección en la que una vez hecho todo el trabajo, los doctorados se ponen a pensar en las personas o, como en mi caso, lo que te ayudó más en los días de frustración, a lo que en parte le debes este gran proyecto. Y reflexionando sobre todo ello, me es imposible no pensar en el CrossFit, que para quien no lo conozca, no es más que un deporte de fuerza y alta intensidad, pero para mí era mi desconexión de todo, mi motivación, mi alegría. . . hasta tal punto que incluso empecé a competir durante este periodo.

Siento que cuando decides empezar un doctorado, debes tener la cabeza bien amueblada para ello. Yo no estaba del todo bien, necesitaba sanar y el deporte fue mi medicina. Además me encontraba en un momento en el que mi vida social brillaba por su ausencia y posteriormente, una pandemia de un innombrable virus llegaba para implantar el teletrabajo en los dos años y medio que de tesis me quedaban. Entonces, comprenderás que mi tesis no se la quiero agradecer a la gente, sino a la terapia que me ayudó a acabar este proyecto siendo mucho más feliz y sana que cuando empecé, el CrossFit.

El deporte fue quien estuvo ahí aguantándome los días malos y devolviéndome la sonrisa. También me ofreció inspiración cuando me sentí atascada con un problema y por supuesto me presentó a los nuevos amigos que tuve que hacer al mudarme de ciudad con el trabajo en remoto y son ellos los hoy día me acompañan.

No pienso poner nombres, porque sé desde el fondo de mi corazón que las personas que me importan ya lo saben y que el 90% de ellas nunca leerán ésto porque no entenderían ni ‘papa’ de un sólo diagrama o ecuación y yo, yo los voy a querer igual. Sin embargo, no me perdonaría cerrar esta sección sin una excepción, o más bien, cuatro.

Mis tutores, Antonio y Pablo, ellos son los que confiaron en mí desde el minuto uno, los que me han enseñado gran parte de las cosas que sé y las que no, ellos me aportaron las herramientas para yo poder volar sola. Fueron los únicos que se leyeron mis trabajos, que analizaron mis resultados, que me guiaron y me aconsejaron. Y para mí lo más importante es que me dieron tiempo desde el principio y me dejaron marcar el ritmo que necesitaba. Es a ellos a quienes les tengo que agradecer esta tesis. Porque el tiempo es lo más valioso que hay en esta vida, y todo el tiempo que ellos me han dedicado a mí no tiene palabras suficientes para que yo lo pueda agradecer. GRACIAS, a los dos.

Y la otra excepción que necesitaba hacer, quizás te la imagines por la dedicatoria. En los 28 años de edad que tengo, la vida me ha enseñado que hay dos personitas que jamás te van a fallar, pase lo que pase. Son mi razón de ser y aunque les costase aceptar la idea de seguir estudiando ‘tropecientos’ años más al principio del doctorado, necesito decirles y dejar escrito aquí un enorme GRACIAS a los dos. Ellos siempre me apoyarán y el calor que he recibido de su parte tampoco tiene manera de ser agradecido.

La vida me tiene preparada una nueva etapa, y después de escribir esta tesis, me siento con ganas de comerme el mundo. Así que agradezco a la vida que me pusiera en mi camino esta gran piedra que tanto he peleado, porque somos el resultado de lo que vivimos, y esta tesis ya es parte de mí.

PUBLISHED AND SUBMITTED CONTENT

The following section describes the papers that have been published or not and that are included as part of the current thesis. The material from this source included in this thesis is not singled out with typographic means and references.

0.1. Preprints

1. Aguilera, A. C., Artés-Rodríguez, A., Pérez-Cruz, F., Olmos, P. M. (2020). Robust Sampling in Deep Learning. arXiv preprint arXiv:2006.02734. [pdf] - It is wholly included in Chapter 5.

0.2. Journals

1. Lopez-Castroman, J., Abad-Tortosa, D., Aguilera, A. C., Courtet, P., Barrigón, M. L., Artés, A., Baca-García, E. (2021). Psychiatric profiles of eHealth users evaluated using data mining techniques: cohort study. JMIR mental health, 8(1), e17116 [pdf] - It is wholly included in Chapter 4.

2. Porras-Segovia, A., Cobo, A., Díaz-Oliván, I., Artés-Rodríguez, A., Berrouiguet, S., Lopez-Castroman, J., ... Baca-Garcia, E. (2021). Disturbed sleep as a clinical marker of wish to die: a smartphone monitoring study over three months of observation. Journal of affective disorders, 286, 330-337. [pdf] - It is wholly included in Chapter 4.

3. Cobo, A., Porras-Segovia, A., Pérez-Rodríguez, M. M., Artés-Rodríguez, A., Barrigón, M. L., Courtet, P., Baca-García, E. (2021). Patients at high risk of suicide before and during a COVID-19 lockdown: ecological momentary assessment study. BJPsych open, 7(3). [pdf] - It is wholly included in Chapter 4

0.3. Journals accepted for publication

1. Aguilera, A. C., Olmos, P. M., Artés-Rodríguez, A., Pérez-Cruz, F. (2021). Regularizing Transformers With Deep Probabilistic Layers. arXiv preprint arXiv:2108.10764. [pdf] - It is wholly included in Chapter 2.

CONTENTS

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
APS	Average Precision Score
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
CGI	Clinical Global Impression
C-GMVAE	Conditional Gaussian Mixture Variational Autoencoder
CNAQ	Council on Nutrition Appetite Questionnaire
CNN	Convolutional Neural Network
CSSRS	Columbia Suicide Severity Rating Scale
DGM	Deep Generative models
DSM	Diagnostic and Statistical Manual of Mental Disorders
EHR	Electronic Health Record
ELBO	Evidence Lower Bound
EMA	Ecological Momentary Assessment
FFL	Feed-forward Layer
FN	False Negative
FP	False Positive
GAN	Generative Adversarial Networks
GLUE	General Language Understanding Evaluation
GMVAE	Gaussian Mixture Variational Autoencoder
GNN	Graph Neural Networks
IBP	Indian Buffet Process
ICD	International Statistical Classification of Diseases and Related Health Problems
IDS	Inventory of Depressive Symptomatology
ISI	Insomnia Severity Index
LM	Language Model
LSTM	Long Short-Term Memory

MBTI	Myers-Briggs Type Indicator
ML	Machine Learning
MLE	Maximum Likelihood Estimation
MLM	Masked Language Modeling
MLP	Multi Layer Perceptron
MoG	Mixture of Gaussians
NLP	Natural Language Processing
NLU	Natural Language Understanding
NN	Neural Network
NoRBERT	Noisy Regularized Bidirectional Encoder Representations from Transformers
NPV	Probability of non-disease given negative test result
PPV	Probability of disease given positive test result
PVR-E	Probabilistic Variance Reducer per Epoch
PVR-M	Probabilistic Variance Reducer per Mini-batch
ReLU	Rectified Linear Units
RNN	Recurrent Neural Network
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SD	Standard Deviation
SGD	Stochastic Gradient Descent
SI	Suicide Ideation
SNLI	Stanford Natural Language Inference
SPFM	Sparse Poisson Factorization Model
SST	Stanford Sentiment Treebank
STB	Suicidal Thoughts and Behaviours
TN	True Negative
TP	True Positive
TREC	Text Retrieval Conference Question Classification
VAE	Variational Autoencoder
VR-E	Variance Reducer per Epoch
VR-M	Variance Reducer per Mini-batch

1. INTRODUCTION

1.1. Motivation

Imagine you are told that no doctor is available to attend your needs and analyse your current state provoked by some kind of illness. However, a computer will diagnose you and send you your treatment. How would you react? Would you trust in this kind of clinical robot? Maybe you are shocked and disagree in first instance, but the truth is that we are not far away from this situation nowadays.

We are living the Artificial Intelligence (AI) era. Whatever we want, we are surrounded by technology, computers, intelligent devices and incredible machines able to perform more and more complex activities from our daily routine. We have built, in a scarce decade, a parallel universe fed by data, data and more data. In consequence, the quantity of information stored in what we call now 'the cloud' really scares. We are not aware of the power of these stored numbers. They, with the help of **Machine Learning** (ML) [1], can be converted into weapons capable of destroying our human rights and integrity or, in the contrary, save our lives or help us to survive in an easier and more comfortable way.

Think about your mobile phone telling you what is the best road to take to arrive faster to your job, or the advertisement about that product your needed that suddenly appeared in the perfect moment to buy it. What about the board predicting what your wanted to write or the facial recognition to unlock your personal device? If you were asked again the same question, would you trust in a robot to diagnose you? Analysing all that AI has achieved until now, you may see it in a different way.

This thesis is delimited in the scope of ML, a multidisciplinary field of which unifying thread is the transformation of sets of data into information. This is reached through problems resolution such as feature extraction, classification, regression or, what constitutes the task with more complexity, the determination of probability distributions. In this last approach it is particularly interesting the denominated **Bayesian or probabilistic modeling**, where this thesis will be based. It allows in a very intuitive and direct way the incorporation of our previous knowledge, quantify the uncertainty and generate artificial examples with the same properties as the input samples. This easy explainable characteristic will be one of the keys to apply this kind of models to health, the main application of the present work. However, before going into detail about the problem solving and the application of our researching, let's see a little bit more about the tools we are going to use.

Recently, probabilistic modeling has been included in architectures from **deep learning** [2], [3], giving rise to methods as Generative Adversarial Networks (GANs) [4]–[6] or Variational Auto-Encoders (VAEs) [7], with a great ability to carry out the estimation of densities in high-dimensional spaces through the determination of a latent projection

in the observations. However, the Bayesian approach goes beyond, modeling the uncertainty in the determination of parameters from deep neural networks and providing a regularization strategy which allows to fight the bad conditioning of this problem. It is defined Bayesian deep learning [8] and corresponds an alternative methodology to non probabilistic techniques in the regularized training of deep networks.

Inside DL world, we know as **Transformers** [9] those models applied to **Natural Language Processing** (NLP), one of the objective scenarios to study in detail during this thesis. In the most complex and unstructured data bases, NLP can serve as a starting point, as for example in the Electronic Health Record (EHR) from a hospital patient, where any kind of data is combined with a great proportion of fields in a free text form. Because of the many applications we can find from the NLP community, studies in this technology are growing with non-stop.

Our development of algorithms based on probabilistic models in combination with Transformers will be focused in obtaining solutions for data exploration in health, an application with a huge social relevance, and more precisely, we will face the characterization of human behaviour in population with mental disorders. From all data resulting due to digital interaction on the patient side, defined generically as **digital phenotype** [10], the principal goal is to obtain useful, explainable and unbiased information for the medical staff, carers and family of the patients. All of this has the aim of improving the care cycle of patients from the psychiatry branch in medicine. This data is composed of heterogeneous values such as questionnaires with medical validity, diagnoses, treatments, social situation, age, the doctor evaluation, indicators collected by wearable devices about physical activity or sleeping, or information from the personal mobile phone with a important component about social activity. Even though we will not analyse the whole phenotype during this thesis, we will make use of a great proportion and show the advantages and utilities we are capable of developing from them.

1.2. Contributions

We present a heterogeneous thesis dealing with the regularization of deep learning models and data exploration in health thanks to the incorporation of the Bayesian approach and Language Models (LM). We contribute to the ML community in very different branches but with a very define common thread. Throughout this work, we mainly address the following two branches:

1. Regularization of deep learning models such as Convolutional Neural Networks (CNNs) or Transformers.

2. Application of ML tools in the diagnosis of mental disorders.

Moreover, the contributions of the thesis are collected in a series of papers, mentioned at the beginning of the document. We reference 3 clinical papers and one technical (accepted

with minor corrections) published in journals. We do also have one preprint in arXiv and two other papers in process of writing.

1.2.1. Regularization of CNNs through robust sampling

Regarding the regularization line of research, we show the results of a smart mechanism for selecting the samples from the minibatch during the training of CNNs in an image classification problem. We present the experiments in well studied networks, VGG [3] and All-CNN-C [11], as a good comparison in the literature. We base our idea in a theoretical work [12] about the variance reduction in the real risk and reproduce it through the repetition of the worst classified samples. This work serve as a touchdown with deep learning and regularization models we will face during the thesis.

1.2.2. Regularization of LMs by a probabilistic layer based in a VAE

This is the main work during the thesis. It is composed of a complete study of our regularization technique based in probabilistic models throughout the main LMs from NLP, from sequence to sequence architectures [13], [14] until Transformer-based models such as BERT, RoBERTa and XLM-R [15]–[17]. To conclude, it is closed with an external application for data augmentation to be applied in other NLP tasks.

The proposal idea is based in a structured noise injection to the embedding of the training sentences at some point of the LM structure. This is achieved thanks to a Gaussian Mixture Variational Auto-Encoder (GMVAE) [18], which reconstructs the embeddings with a forward and backward step into the projection to a hidden space. This reconstruction is responsible of some noise injection following the probability distributions defined in the architecture, provoking a regularization effect in the networks where it is applied.

In seq-to-seq models, with and without attention, we reconstruct missing words in some text corpora with a more diverse topic constitution. However, in Transformers, we obtain this same result or not depending on the layer depth we regularize. That is, we do only regularize one layer and present Deep and Top NoRBERT, when including the GMVAE block in the deeper layers or precisely before the top classification layer respectively. Top NoRBERT is also able to find a topic-diverse generation of text, improved to longer sequences regarding seq-to-seq structures. On the contrary, Deep NoRBERT gets better score when reconstructing a sentence to its original form but in any case, we get a regularizer effect capable of delaying the overfitting point. We validate our experiments with different Transformers models and datasets.

Finally, related to this work, we present Contextual NoRBERT, with a conditioned GMVAE including information from neighbors words on each embedding reconstruction. We present improvements in the classification score of different datasets in both vanilla and contextual NoRBERT.

1.2.3. Applying Transformers in mental health diagnosis.

Following the analysis of Transformers, we present PsyBERT based in BERT [15] architecture and inspired by BEHRT [19]. It consists on a Transformer Encoder with a modified construction of the input embedding adapted to work with sequential data from the EHR as it was natural language. We create several dictionaries, one per each feature to be included from the EHR and concatenate the embeddings. Moreover, we include an additional embedding from a free text field combining information from all visits in a patient. AS a result, we obtain a sequential model capable of deal with heterogeneous data such as coded diagnoses, age, text or sex.

We apply this architecture in solving two different tasks. On the one hand, we impute the missing diagnoses in the EHR from a psychiatric hospital in Madrid. We face with very high missing rates, more that the half of data, comorbidity and a total number of 768 different diseases. We present the results validated by an expert in psychiatry sector. On the other hand, we deal with the diagnosis of delusional patients, a very low prevalence disorder and very difficult to be recognize in many situations. Again, we validate the results with the doctors and compare different models in the analysis of the experiments.

1.2.4. Applying probabilistic methods in mental e-health questionnaires.

The last line of research from the mental health perspective consist in a feature extractor model applied to different questionnaires in order to find different behaviour patterns in the users. These users correspond to psychiatric patients, so we combine these results with the EHRs in order to find common patterns among patients from the same group of disorders. The before mentioned feature extractor is a probabilistic model based on the Indian Buffet Process (IBP) [20], [21]. More exactly, it consists on a nonparametric latent feature model that proposes a sparse decomposition of the variables and is called Sparse Poisson Factorization Model (SPFM) [22].

We present three different works following this procedure. Firstly, we applied the methodology in a general health questionnaire in order to find different profiles among the patients. We related these profiles with the diagnoses of the patients and obtain different conclusions about the mental health of the users regarding the punctuation on sets of questions. Secondly, we enclosed the problem into the search of relation between sleep alterations and suicidal thoughts. During this work with show the correlation between these two variables and conclude the results within a time window procedure o incorporate sequential information. Thirdly, we conclude this line of works with a last application during covid-19 lockdown. We make use of the same questionnaire as in the previous work and use also time information in order to analyse the suicidal risk in patients before a during lockdown. We conclude a decrease in the suicide risk during the lockdown, all the contrary of what we could have expected.

In all these studies, we validated and worked in collaboration with doctors from the

psychiatry sector. We always used the data that they considered more valuable and we served as a platform between the model output and the expert interpretation.

1.3. Organization and connections

The following points define very shortly the work collected on each of the main chapters in this document. As it is a heterogeneous thesis, we present the results in the easiest readable way, what does not have to coincide with the real chronological sequence of the studies development but they are sorted by importance value:

1. **Chapter 2**: Enrich the generation of missing words in a text corpora with a novel regularization technique via Variational Auto-Encoders combined with LMs.
2. **Chapter 3**: Study Transformers networks in the diagnosis of mental disorders with heterogeneous data from the EHR.
3. **Chapter 4**: Relate mental disorders with human behaviour patterns according to a probabilistic model based in matrix factorization for data exploration from questionnaires.
4. **Chapter 5**: Regularize other deep network models by a smart batch selection based in variance reduction.

While in Chapter 3 we will use the EHR (Electronic Health Record) in order to predict undetected or missing diagnoses from patients with any mental disorder, in Chapter 4 we will study psychiatric patterns regarding e-health questionnaires information. All our medical studies are placed in the field of psychiatry and we use different patient sources of information which are used in the often practice by the expertise professional. Therefore, our tools can serve as an additional resource they can exploit in prognosis. We focus our work in the diagnosis-related problems. With both chapters studies, we will find connections between behavioral disorders, detect patterns in patients' profiles, correct errors in the EHR, fill missing information or treat comorbidity.

Regarding Chapters 2 and 5, we study two different techniques of regularization. The first one, also considered the most important contribution from this thesis, is applied to language models and the second is used in computer vision.

Finally, Chapters 2 and 4 are joined by models with a Bayesian approach. In the former the regularizer is based in a Variatonal Auto-Encoder while the latter consists in several applications of a non-parametric probabilistic method.

2. NOISY REGULARIZED LANGUAGE MODELS

2.1. Introduction

Language models (LM) have grown with non-stop in the last decade, from sequence-to-sequence (seq2seq) architectures to attention-based Transformers. However, regularization is not deeply studied in those structures. In this work, we use a Gaussian Mixture Variational Autoencoder (GMVAE) as a regularizer layer. We study its advantages regarding the depth where it is placed and prove its effectiveness in several scenarios. Experimental result demonstrates that the inclusion of deep generative models within Transformer-based architectures such as BERT, RoBERTa or XLM-R can bring more versatile models, able to generalize better and achieve improved imputation score in tasks such as *SST-2* and *TREC* or even impute missing/noisy words with richer text.

Deep Generative Models (DGMs) have become a cornerstone in modern machine learning due to their ability to learn abstract features from high-dimensional spaces to generate new data ([4], [23]). In the field of Natural Language Understanding (NLU), state-of-the-art is dominated by attention-based probabilistic models, a class of explicit DGMs that can be trained with Maximum Likelihood Estimation (MLE) approaches [24].

Regarding other well known DGMs such as Generative Adversarial Networks or GANs [4], so far for NLU they have not shown the same outstanding results that they achieve for image processing ([6], [25], [26], [27]), mostly due to the discrete nature of the data, which leads to non-differentiable issues, mode collapse and optimization instability ([28], [24]). To tackle these and other issues, recent contributions propose the use of Reinforcement Learning techniques to optimize the GAN loss function ([29], [30], [31], [32]), continuous approximations to discrete sampling ([33], [34], [35]), learning a low-dimensional representation through autoencoders ([36], [37], [38], [39], [40], [41], [42]), use other approaches ([43], [44]) or even more recently combine them with transformers [45]. Besides, explicit DGMs such as variational autoencoders (VAEs) have also been proposed in several NLU approaches again with limited success ([46], [47], [48], [49], [50], [51], [52]). Some of the pioneers in this field were [53], who proposes a RNN-based VAE for text generation. Even in an extent, [54] combine a VAE with a discriminator to build a hybrid model that solves the text generation problem. In all these works, both GANs and VAEs are at the core of the NLU model, and hence are fully responsible to capture the semantic structure and generate text. For this particular task, they are still not competitive with attention-based probabilistic models [24].

In this work, we propose to exploit DGMs for NLU in a completely novel and different way. Instead of training a DGM to solve a NLU task, we rely on a hybrid model in which a transformer-based architecture like BERT [15] is combined with a VAE, which is placed inside its structure as a stochastic layer that helps to learn a richer hidden space, enforcing

a regularization effect. In particular, we use a hierarchical VAE that implements a mixture of Gaussians in the latent space (GMVAE) [18], since it is able to capture more complex data in an easier way than the traditional vanilla VAE. In a similar way, [55] and [56] built fusion models taking advantage of a pre-training process as we explain later. Nevertheless, they only focused on a basic seq2seq architecture.

Regularization in deep learning has risen up from the beginning of Neural Networks (NN) with the extensively use of tools such as dropout [57], early stopping [58], data augmentation [59], weight decay [60], or more recently transformer-based dropout, DropAttention [61], which helps models to generalize. However, regularization in NLUs is a much-less explored field and none of these tools experience the same versatility as our proposal in this paper, in which the GMVAE performs a controlled and structured noise injection within the NLU deep network. When combined with BERT, we name our model as NoRBERT (Noisy Regularized BERT) and we conclude that the effect of the stochastic layer is very different depending on the transformer layer where it is placed. If the layer is placed at the bottom of the structure, it improves BLEU (Bilingual Evaluation Understudy) score [62], what coincides with the goal of traditional regularization mechanisms and GLUE (General Language Understanding Evaluation) benchmarks. On the contrary, when placed at the top, it drives more versatile topics when imputing missing words.

Mainly, we illustrate our approach in word imputation problems (masking tokens in the source text corpora) using a BERT transformer network or any of its variants. However, we extend those results in other scenarios as a prove of our methodology effectiveness. The contributions of our work, sorted by importance grade, are as follows:

- We demonstrate gains in a setup of pretraining masked language modeling (MLM) by better BLUE score in a large set of examples through Deep NoRBERT.
- We also show the versatility of the method to impute missing words. We used Top NoRBERT for this result.
- Then, we apply our model as a data augmentation tool to solve a classification task with improvements over the baseline without any data augmentation at all.
- As the first step, we explore the GMVAE regularization effect in traditional seq2seq models with and without attention mechanisms and expose its results as a positive evidence to explore larger models.
- At the begining of the chapter, we include the explanation of the regularization functionality in a simple well-known problem which is the classification of Fashion MNIST images and compare it with dropout mechanism.

The code to generate our results in FMNIST is available in an open repository[1] and the code for NoRBERT model is in another one[2].

[1] https://github.com/AuroraCoboAguilera/NoRClassifier
[2] https://github.com/AuroraCoboAguilera/NoRBERT

This work is organized as follows. Firstly, in Section 2.2 we describe some related work which is key to understand the paper: the VAE, and more precisely the GMVAE, as the main structure of the regularizer and transformer networks with BERT as our main model to be studied. In addition, throughout the paper we include some experiments with RoBERTa and XLM-R as improved versions of BERT but with few differences in their structures. Secondly, in Section 2.3 we explore a basic example of applying our idea in the classification of the Fashion MNIST dataset. This is a useful prove of the stochastic layer effect and its effectiveness in a completely different scenario, comparing it with dropout. Thirdly, in Sections 2.4 and 2.5 we describe in detail our model. We start we Noisy Regularized seq2seq and finish with NoRBERT and two variants of it, Top and Deep NoRBERT, depending on the transformer layers where we apply the regularization. Then, we present the results of these two options in Section 2.6. Finally, we study our model as a data augmentation tool in a classification set-up in Section 2.3. As a conclusion, in Section 2.8 we resume our contributions and mention some future lines of research.

We wanted to focus the thesis in this main work, so this is our more extensive piece, where we have given more effort during these 4 years of doctorate.

2.2. Related work

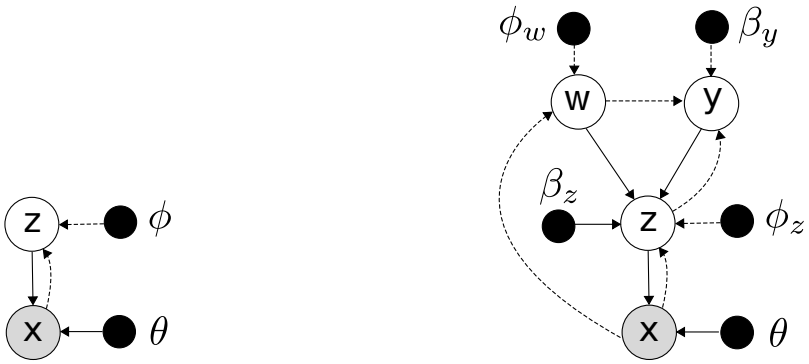
2.2.1. Variational Autoencoders with Gaussian mixture priors

A VAE [23] is a class of density estimator that consists on two networks, an encoder and a decoder or generator, that builds a regular latent space with the help of probability distributions. The properties of the organized latent space allow not only the reconstruction of the input data but also the generation of new instances from a sampling procedure. In a standard vanilla VAE, see Figure 2.1a, the low-dimensional latent space follows a Gaussian prior distribution, i.e. its parameters, the mean and covariance matrix of $p(x|z)$, are parameterized through the decoder network with input z. Variational inference of the model parameters is achieved by maximizing a lower bound on $\log p(x)$, which in turn depends on a flexible NN parameterized distribution $q(z|x)$ that approximates the true posterior $p(z|x)$:

$$\mathcal{L}_{ELBO}(\theta, \phi, x) = \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \mathcal{KL}\Big[q_\phi(z|x)\,\|p(z)\Big], \tag{2.1}$$

where $\mathcal{KL}(q|p)$ is the KL divergence between distributions q and p and acts as a regularization in the evidence lower bound (ELBO) objective. The graphical model of the variational approximation, $q(z|x)$, is indicated in Figure 2.1a with dotted lines.

The flexibility of VAEs has encouraged the study of different priors and architectures to obtain models capable of inferring more complex structured data. That is the case of using a Mixture of Gaussians (MoG) as the prior distribution $p(z)$ for the latent space because it helps to capture the multimodal nature of some data [18], [63]. We refer to this method as GMVAE, and its graphical model is shown in Figure 2.1b. The generative model of the



(a) VAE. (b) GMVAE.

Figure 2.1: The directed graphical models into consideration. Solid lines denote the generative model and dashed lines the variational approximation. The shaded variables are considered the observed inputs, the dark units are the networks parameters to be optimized and the units that are left are the latent variables.

GMVAE proposed by [18] is characterized by the following distributions:

$$p(z) = \int p(z|w,y) \cdot p(w) \cdot p(y) dw dy \tag{2.2a}$$

$$p(w) = \mathcal{N}(0, \mathcal{I}) \tag{2.2b}$$

$$p(y) = \text{Mult}(\pi), \quad \pi_i = \frac{1}{K} \tag{2.2c}$$

$$p_{\beta_z}(z|w,y) = \prod_{k=1}^{K} \mathcal{N}(\mu_{\beta_{y_k}}(w), \Sigma_{\beta_{y_k}}(w))^{y_k==1} \tag{2.2d}$$

$$p_\theta(x|z) = \mathcal{N}(\mu_\theta(z), \sigma\mathcal{I}), \tag{2.2e}$$

where $\mu_{\beta_{y_k}}$, $\Sigma_{\beta_{y_k}}$ and μ_θ are neural networks. $\mu_{\beta_{y_k}}$ and $\Sigma_{\beta_{y_k}}$ indicate a different NN per component in the MoG and K is the total number of components. The posterior distribution of z, w and y given x is chosen according to the following factorization

$$q_{\phi_z}(z|x) = \mathcal{N}(\mu_{\phi_z}(x), \Sigma_{\phi_z}(x)) \tag{2.3a}$$

$$q_{\phi_w}(w|x) = \mathcal{N}(\mu_{\phi_w}(x), \Sigma_{\phi_w}(x)) \tag{2.3b}$$

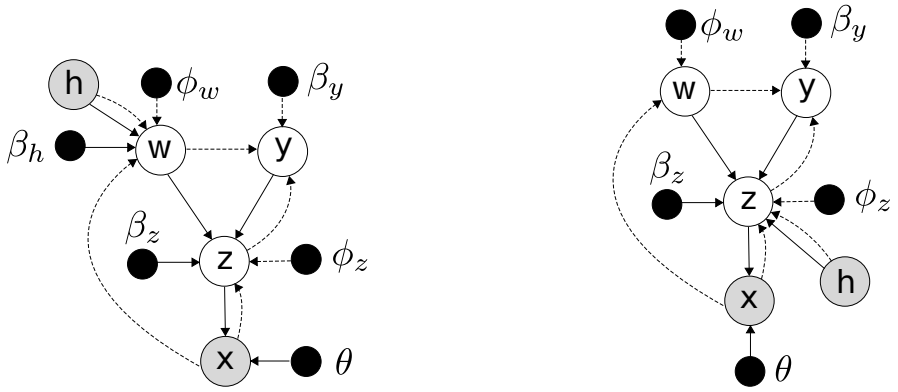
$$q_{\beta_y}(y_j == 1|w,z) = \frac{p(y_j == 1) \cdot p_{\beta_z}(z|y_j = 1, w)}{\sum_{k=1}^{K} p(y_k == 1) \cdot p_{\beta_z}(z|y_k = 1, w)}, \tag{2.3c}$$

where again μ_{ϕ_z}, Σ_{ϕ_z}, μ_{ϕ_w}, and Σ_{ϕ_w} are dense neural networks, resulting in the following ELBO:

$$\begin{aligned} &\mathcal{L}_{ELBO}(\theta, \phi, x) = \\ &\mathbb{E}_{z\sim q_{\phi_z}}\left[\log p_\theta(x|z)\right] - \mathbb{E}_{w\sim q_{\phi_w},\, y\sim p_{\beta_y}}\left[\mathcal{KL}\left[q_{\phi_z}(z|x)\,\|p_{\beta_z}(z|w,y)\right]\right] - \\ &\mathbb{E}_{z\sim q_{\phi_z},\, w\sim q_{\phi_w}}\left[\mathcal{KL}\left[p_{\beta_y}(y|w,z)\,\|p(y)\right]\right] - \mathcal{KL}\left[q_{\phi_w}(w|x)\,\|p(w)\right] \end{aligned} \tag{2.4}$$

The Conditional GMVAE

When we deal with conditional DGM, we mean that the entire generative process is conditioned on some extra observed inputs. [64] presented Conditional Variational Autoencoder (CVAE), where the observations modulate the Gaussian prior. In a similar way, we have studied two architectures to condition our distributions on an input that we have defined h. In this section, we expose the changes applied and describe the two versions of C-GMVAE that we have explored, referred as models A and B.



(a) C-GMVAE model A.

(b) C-GMVAE model B.

Figure 2.2: The directed graphical models considered for the C-GMVAE in the work. Solid lines denote the generative model and dashed lines the variational approximation.

The first architecture (model A) that we tried is shown in Figure 2.2a. For its implementation we had to change the prior distribution of w as $p(w|h) \sim \mathcal{N}(\mu(h), \Sigma(h))$, where the mean and variance of the normal distribution are parameterized by dense nets, and $q_{\phi_w}(w|x,h) = \mathcal{N}(\mu_{\phi_w}([x;h]), \Sigma_{\phi_w}([x;h]))$, where we only concatenate h to the original input x. The main drawback of this model is that the reconstruction as we performed it (compute z from x through the inference model and then reconstruct x from this z by the generative model) does not use the observed h.

The graph in Figure 2.2b belongs to our second version (model B), the one we applied to the presented results. In contrast, in the generative model, we now maintain the original prior of w but condition the z distribution on h as Equation 2.5a. For the variational family, we modify the encoding of z as Equation 2.5b.

$$p_{\beta_z}(z|w,y,h) = \prod_{k=1}^{K} \mathcal{N}(\mu_{\beta_z}([w;h]), \Sigma_{\beta_z}([w;h]))^{y_k==1} \tag{2.5a}$$

$$q_{\phi_z}(z|x,h) = \prod_{k=1}^{K} \mathcal{N}(\mu_{\phi_z}([x;h]), \Sigma_{\phi_z}([x;h])) \tag{2.5b}$$

2.2.2. Language Models

In this section, we recap the language models that are regularized during the study. Then, we start with basic sequence to sequence models to finish with Transformers.

Sequence to sequence

In the field of NLU, architectures based in an encoder-decoder framework have been considered a milestone and were constantly studied in the literature until the appearance of Transformer-based pre-trained models. That is the reason why we decided to start to research in the idea of probabilistic regularization in them, as a baseline, and then, at the same time I enriched my knowledge in the field, I could go into a more complex scenario within Transformers.

A sequence to sequence (seq2seq) model is an architecture composed of an encoder which maps the input sentence into a fixed-size vector and a decoder to map this vector into a target sentence. This sequence to sequence, or sentence to sentence, transformation was what made them successed as a machine translation tool [65].

Firstly, those that rely on RNNs generate a sequence of hidden states, each h_t as a function of the previous one, h_{t-1} [65]. However, they show limitations for long sentences since they encode the semantic and syntactic information of a whole sentence in a single vector. Then, LSTM-based units appeared to give better results [65]. Additionally, attention mechanisms [13], [14] allow this kind of models to focus on the relevant parts of the source sentence, acting as an alignment system between encoder and decoder and improving the performance.

In this work, we will study both perspectives, with and without attention, and they will serve as a precedent to the final NoRBERT construction.

Transformer networks

Over the last couple of years, Transformers [9] have become a revolution in the field of NLU ([66], [67], [68], [69], [70]) due to their ability to capture longer-range linguistic structure. Unlike previous works ([65], [13], [14]), they rely entirely on self-attention to compute the latent representations of the sentences.

Transformer-based models are usually applied in a transfer learning perspective ([15], [71], [72], [73], [74], [75], [76], [77], [78], [79], [80], [81] [82] [83]) that allows users to train smaller datasets in a specific task quicker and more accurate than doing it from scratch. Firstly, you need a pre-trained model that has learned contextualized text representations in a general unsupervised scenario with a large text corpus. Afterwards, you can fine-tune the model using a small database with the addition of few parameters or layers in a downstream task. This is the case of BERT (Bidirectional Encoder Representations from

Transformers) [15], providing a pre-trained Transformer text encoder as a general LM for any downstream task. It generally learns bidirectional representations from unlabeled text by the conditioning on right and left context information. Since its appearance, several BERT-based models have emerged ([16], [84], [85]) and today they dominate the leaderboard[3] in GLUE benchmarks [86]. In this work, we will use different versions of BERT. In general terms, RoBERTa [16] makes a different choice in the pretraining hyperparameters and XLM-R [17] uses a dataset with samples in different languages. We will see experiments for all three models in order to prove the effectiveness of our method in a larger range of NLU problems.

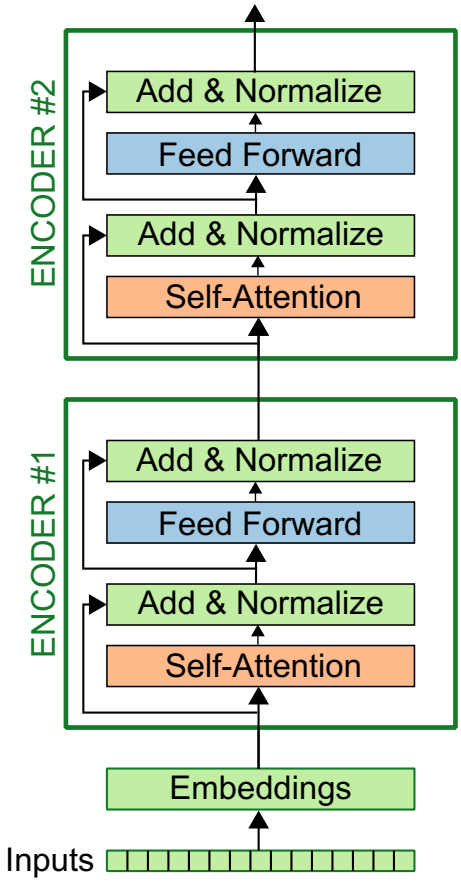


Figure 2.3: Diagram of BERT structure with two encoder layers and no task-specific layer.

Figure 2.3 shows a diagram with the structure of BERT for the first two layers. Basically, it is composed of a first step with the computation of the input sentences embeddings, then a pile of transformer encoder layers, and then, if it is necessary we can apply any task-specific layer on top. Each of these encoder layers consists on two blocks, a multi-head self-attention mechanism and a feed forward network, both with a normalization following them.

Regarding the implicit regularization mechanisms within BERT, dropout and weight decay are applied through all the structure: in the fully connected layers in the embeddings,

[3] https://gluebenchmark.com/leaderboard

encoder, pooler and in the attention probabilities with rates of 0.1 and 0.01 respectively (in the pretrained version we use, see Section 2.6.2). Furthermore, our BERT pretraining follows the MLM, a strategy that may work as a regularizer at some extend since it hides and corrupts some tokens from the original sentences.

BERT makes use of WordPiece embbedings [87] with a vocabulary size of 30000 tokens in the base models we use. They are pre-trained in the datasets of Book Corpus [88] with 800M words and English Wikipedia with 2500M words.

2.3. GMVAE as a regularizer in deep neural networks

In this work, we put forward GMVAEs as a robust stochastic layer to enforce regularization in a deep NN, with particular focus on NLU and Transformers. Before describing the methodology in a complex LM, we want to illustrate our approach in a simpler setup, in which we regularize a deep six-layer MLP over the Fashion MNIST (FMNIST) database[4] [89].

The dataset is composed of 28x28 images in grayscale associated with a label from 10 classes. We divide the set in 12000 samples for training and 48000 for validation. The test set has 10000 images. The only preprocessing step is the normalization to 0.5 mean and variance.

The model used for these experiments consists on 9 linear layers with RELU as the activation function. The size of the output features on each layer is, from bottom to top, 700, 600, 512, 256, 128, 64, 32, 16 and 10, which corresponds with the number of classes. We employ a negative log-likelihood loss function and the SGD with a learning rate of 0.01 for the optimization.

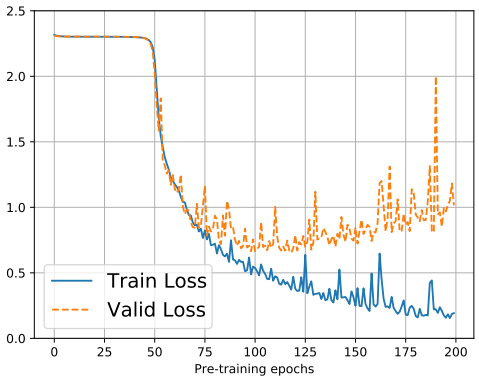


Figure 2.4: Pre-training a six layer MLP for FMNIST.

In Figure 2.4 we show the train/validation cross entropy loss of the NN in a completely unregularized training (no dropout or weight decay whatsoever). Validation error begins

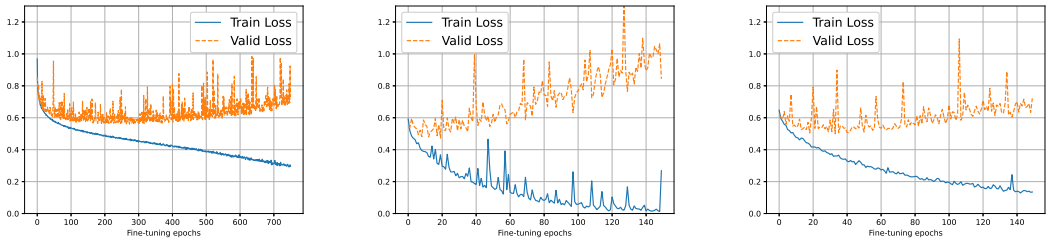
[4] https://github.com/zalandoresearch/fashion-mnist

to raise up from epoch 120. Now we perform the following experiment. We get the NN parameters at epoch 119, at which overfitting was not yet noticeable (i.e. early stopping procedure), and we introduce two types of regularization layers between the first two MLP layers:

1. A standard dropout layer with erase probability p.
2. A GMVAE layer trained using the 700-dimensional internal representation of the first MLP layer. For every output from the first MLP layer, the GMVAE layer first computes a latent low-dimensional representation sampling from the GMVAE posterior distribution in (2.3a)-(2.3c) to then provide at the output a reconstruction sampled from the generative model in (2.2a)-(2.2e).

Note that the GMVAE layer, as dropout, is introducing a certain level of distortion over the input vector but, unlike dropout, such distortion is not independent to the input vector, as for some atypical vectors the reconstruction noise will be larger. This allows the network to explore diverse regions at the input of the following layer. In Figure 2.5a we show the train/validation cross entropy loss when the layer 1 parameters are frozen (so the GMVAE input distribution is not changing) and we keep training MLP layers 2-6. In Figure 2.5 we also show the performance when dropout with $p = 0.1$ (b) and $p = 0.5$ (c) is used instead of the GMVAE layer.

On the one hand, observe the inability of dropout to compensate the overfitting of the network. On the other hand, due to the controlled noise injection, the GMVAE avoids overfitting even after an excess of additional epochs. With these figures we can state that the training loss decays much more slowly in our model with a score of 0.3 after 700 epochs, while in the dropout case it drops off almost to zero after 150 epochs.



(a) With GMVAE layer. (b) Dropout probability of 0.1. (c) Dropout probability of 0.5.

Figure 2.5: Fine-tuning a six layer MLP for FMNIST with the GMVAE regularization layer placed after the first MLP layer (a) and with dropout in the first layer ((b) and (c)).

Finally, we have included a graph with the training of the NN when dropout ($p = 0.1$) is applied to every layer, since it is the traditional way this regularization mechanism usually works. As seen in figure 2.6, even with this approach, the overfitting behaviour is still reached in more extend than with the GMVAE.

Figure 2.6: Fine-tuning a six layer MLP for FMNIST with dropout in all layers.

With this example, we simply want to put forward the use of a DGM (a GMVAE in our case) as a potential regularizer with additional flexibility, compared to simpler solutions such as dropout. A detailed cross-validation analysis of what kind of regularization method optimizes the classification performance in this particular setting is not relevant at this point. In the following, we show how the use of GMVAE layers is able to enhance the performance of complex LM such as seq2seq (Section 2.4) and pre-trained networks such as BERT (Section 2.5), which of course have already been trained with its own regularization methods (including dropout).

2.4. Improving seq2seq with GMVAE layers: NoR-seq2seq

The main idea of our work is the integration of a DGM in different language architectures. The objective of this hybrid model is the addition of structured random noise to the suitable hidden vectors of these architectures so we obtain more robust solutions. In this section we define the way of regularizing the seq2seq model, that is, the place where we include the GMVAE layer within the LM architecture and how we train it. Firstly, we describe the regular version [13] and next the one with attention [14]. In the latest, we propose several scenarios. For every case consider the regularizer as a black box (in figures is a box in green color with the GMVAE name), where the input is a hidden vector from the LM, h, and the output its reconstruction by the GMVAE, $\hat{h}$.

The training procedure is always the following:

1. Pre-train the LM model as the original version does.
2. Train the GMVAE with the corresponding hidden vectors regarding the place where it is going to be applied and the training sentences from the dataset.
3. Incorporate the GMVAE layer within the LM structure.
4. Finetune the LM model with the GMVAE layer reconstruction, freezing the parameters in the structure below the regularization.

2.4.1. Sequence to sequence

In this architecture, we propose to train a GMVAE over the encoder output as shown in Figure 2.7. In the fine-tuning step, the encoder is fixed and the decoder is re-trained taking as inputs the GMVAE noisy reconstructed vectors.

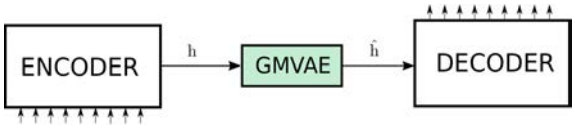
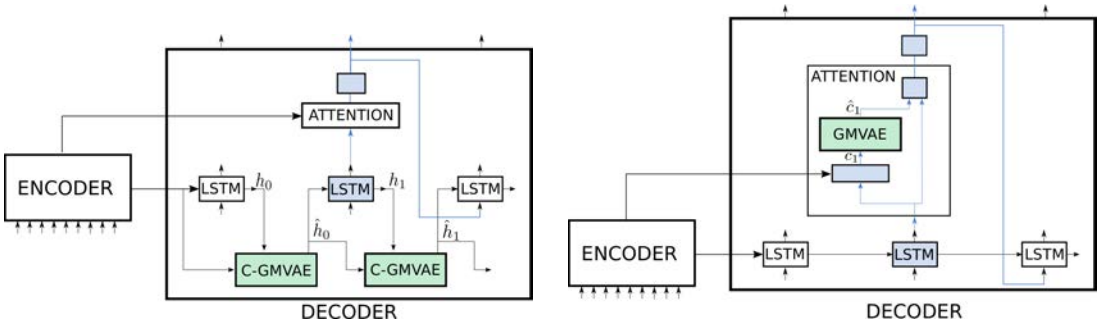


Figure 2.7: Diagram of the regularized seq2seq model.

2.4.2. Sequence to sequence with attention

Now, as the decoder attends to the encoder hidden states at each time step, the previous approach (in 2.4.1) results to be insufficient. In this section we present two kind of methodologies: the option 1 regularizes the hidden states in the decoder LSTMs with a Conditional GMVAE (C-GMVAE), and the option 2 the attention vectors with a GMVAE.

The option 1 aims to regularize the hidden states of the decoder at each time step $(h_0, h_1, ...h_T)$. To achieve this task, we train a C-GMVAE with pairs of consecutive hidden states (h_i, h_{i+1}) from the training sentences, the first one acting as the conditioning input and the second as the input to be reconstructed. See Section 2.2.1 for details on the C-GMVAE. At each time step, the C-GMVAE receives the previous reconstructed state and the current hidden state $(\hat{h}_{i-1}, h_i)$ to reconstruct the latter $(\hat{h}_i)$. As an exception, the first iteration reconstruction, $\hat{h_0}$, is conditioned to the encoder output. Figure 2.8a shows a diagram with this approach. We highlight in blue the process concerning the step $i = 1$ as an example, but it is repeated from the beginning until the moment the end-of-sentence token is generated.



(a) Option 1.

(b) Option 2.

Figure 2.8: Diagrams of the GMVAE regularized seq2seq model with attention.

Option 2 is structurally simpler. We incorporate the noise in a controlled way, avoiding dependencies on previous states. For that, we propose to introduce the GMVAE layer inside the attention mechanism itself. In particular, the GMVAE layer is trained over the

context vectors ($c_0, c_1, ...c_T$). For more details in the attention structure see [13]. The model is shown in Figure 2.8b, where we train the GMVAE with the context vectors of words from the training sentences. We treat each token independently in the GMVAE, since the context vectors usually attend to no more than one or two tokens, thus not requiring a Conditional GMVAE.

Other models

In the autoregressive model of seq2seq with attention, we, firstly, tried training the C-GMVAE from Figure 2.2a to generate each hidden state conditioned on the previous one. These generated states were the inputs for the next LSTM unit. However, it did not work as good as we expected. Consequently, we changed the process in a way that instead of using the DGM to generate samples, we could take advantage of its latent space and reconstruct the original hidden states from the LSTMs.

Inspired by the first model (Section 2.4.1), we also tried to follow the same idea that is presented as option 1 in Section 2.4.2 but conditioning always in the encoder output instead of the previous hidden state, but it did not improve neither the results that we are presenting in this work.

Regarding the option 2, initially we used a simpler approach, regularizing the attention output after concatenating it with the LSTM output and exactly before applying the classification layer that matches the vocabulary size. Here, the training was not successful and the imputed words did not follow grammatical rules as a LM is expected to do. After this, we tried the integration of the GMVAE in a previous step, as it is successfully explained in this work.

2.5. Improving BERT with GMVAE layers: NoRBERT

2.5.1. Overview

The main work of this chapter is the integration of the GMVAE in BERT through NoRBERT. In this hybrid model, the GMVAE layer alters the BERT hidden embeddings in one particular layer through a project-and-reconstruct operation, adding a structured noise to them and hence enforcing a regularization mechanism, the same perspective we saw in Section 2.4. In other words, we try to break the determinism in exchange of more robust solutions. Unlike in other regularization techniques such as dropout, the reconstruction error plus the observation noise (GMVAE noise for short) of the GMVAE will not be uniform across embeddings, since atypical embeddings will suffer from larger GMVAE noise variance. As a result, the network training will rely less on such noisy embeddings, which we show in Section 2.6 is beneficial for the overall performance.

We want to stress the fact that we use BERT as an exemplary case of how a certain

neural language model can be enhanced by the inclusion of GMVAE layers within. Furthermore, in Section 2.4 we showed how to incorporate the same idea in seq2seq language models with attention and later in Section 2.6.2 we will see the same approach applied to RoBERTa and XLM-R, which both are based in BERT architecture and all this section also applies to them. Moving back to BERT, NoRBERT builds upon a pre-trained BERT model, allowing the integration of the GMVAE in an intermediate step. Comparably to Section 2.4, we follow these four main steps:

1. Pre-train BERT with a masked text corpora, i.e. MLM over unlabeled samples.
2. Train a GMVAE over the space of hidden embeddings coming from input sentences using one particular BERT layer.
3. Include the GMVAE layer inside the structure. The GMVAE will be responsible for adding noise in the propagation of the information, as in the GMVAE layer every input vector is projected into a low-dimensional space and reconstructed back by sampling from the generative model.
4. Retrain the model by fine-tuning all layers above the GMVAE one. The layers below the GMVAE one are not altered so we do not modify the embedding space in which the GMVAE was trained on.

Regarding the base BERT model, for the implementation we use the one from [15]. In the training we use the MLM approach as [16], since it is the straightforward strategy to train transformers in word imputation [72].

2.5.2. Methods

In the same way we analysed several scenarios regularizing seq2seq models, in the study of NoRBERT we explore placing the regularizer in different layers from BERT. Firstly, we consider the consequences when the biggest part of BERT is retrained after placing the GMVAE in one of the first and middle layers. This is referred to *Deep NoRBERT*. Secondly, we look into the effect of the GMVAE on top of the transformer encoder, just before the classification layer that computes the vocabulary logits. We refer to this case as *Top NoRBERT*.

Deep NoRBERT consists on a new version of BERT, with a stochastic layer inside a specific intermediate encoder layer, after the self-attention and before the feed-forward block as shown in Figure 2.9(a). In the step 4, we fine-tune the parameters in the structure above the regularizer, that is, the feed-forward block in the same encoder layer and the whole layers that are on top of it. In our experimental results, we demonstrate gains w.r.t. the base BERT model by including only one GMVAE layer. We tried using GMVAE layers within the BERT structure but resulted in negligible goals.

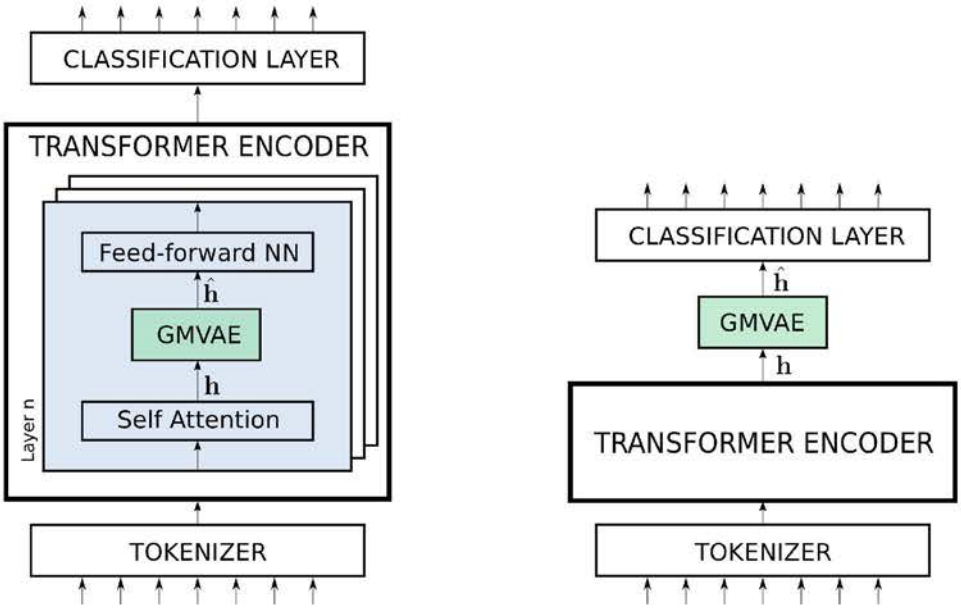


Figure 2.9: Deep NoRBERT (a) and Top NoRBERT (b).

In Top NoRBERT we include the GMVAE stochastic layer on the top of the transformer encoder as represented in Figure 2.9(b). Therefore, the only difference from the original model is that we use a GMVAE to reconstruct the last hidden states before the final token decision. We previously train the GMVAE with the hidden states computed by base BERT for the training sentences. Afterwards, we fine-tune the classification layer of BERT with the stochastic reconstruction integrated.

The main difference in the implementation of both models relapses in the place of the GMVAE within the transformer layer. While in Top NoRBERT we are looking for modifying the upper embeddings in the whole structure, in Deep NoRBERT we are interested in the lower ones. Therefore, we add noise in the output of the feed-forward NN in Top NoRBERT cause it is the closest embedding to the classification layer. Regarding Deep NoRBERT, we also studied to modify the output of the feed-forward but appeared to work worse (in terms of accuracy and BLEU score) so that is why we apply the GMVAE in the output of the self-attention layer.

2.6. Experiments

In first place (Section 2.6.1), we show how the imputation diversity of traditional seq2seq-type models [13] can be enhanced by including a regularizer GMVAE layer inside their structure. We start with a simple seq2seq model and then a seq2seq with attention [14]. Then, in Section 2.6.2, we show the results of regularizing Transformers with the same idea. In this second LM we obtain two different results depending on the variety of NoRBERT used: with Top-NoRBERT we found similar advantages as in seq2seq models but with Deep-NoRBERT instead of augmenting the diversity in the imputation of missing words we improved the BLEU score. We will see then two different advantages our proposal can

achieve depending of the application form.

Data

In Section 2.6.1 we train the models with the *Multi30k* dataset [90]. *Multi30k* is composed of a training set of 29000 sentences and a set of 1000 test sentences with a vocabulary size of 10118 tokens. Although it belongs to a multilingual image description task with its corresponding translations in German, we only focus on the English descriptions. It is not a very large corpora, but we mask some tokens so the scenario gets more complicated to be trained. We use different strategies in the masking process, with higher and lower rates.

In the **first strategy** (for NoR-seq2seq), we use a policy of masked tokens more sophisticated that permits the masking of a less percentage of words but focusing on nouns, verbs, adjectives... That is, ignoring stopwords. We use the English stopwords list from the *nltk*[5] library [91]. We mask the 80% of the sentences and generate two masks in a sentence with a probability of 0.8. Among these, we also generate a third mask with probability of 0.8.

In the **second strategy** (for NoR-seq2seq with attention), we increase the number of masked tokens and do not exclude any grammatical class so any word can be deleted. In this policy, we mask each token with a probability of 0.6. Therefore, we have more [MASK] tokens than proper words.

Afterwards, also in Section 2.6.1 we employ a dataset with more complexity as it is the Stanford Natural Language Inference (*SNLI*) corpus[6] [92]. It has a vocabulary size of 36711 different words. We use the entire preprocessed training data which contains 714667 sentences and a test set of 13350.

In the main experiments with BERT model (Section 2.6.2) we employ *SNLI* with different strategies in the masking process of tokens. The GMVAE is trained for all the tokens of a random set of 50000 training sentences. When saving the hidden states to train the GMVAE a posteriori, we treat each token as an independent input to the GMVAE, ignoring tokens that correspond to padding (they exist due to BERT format of the tokenizer, *WordPiece*).

Later in Section 2.6.2, we deploy a set of experiments with Deep NoRBERT and the three previously mentioned transformers models (BERT, RoBERTa and XLM-R) for which we use three different datasets, *Multi30k*, *SST-2* and *TREC*. Stanford Sentiment Treebank (*SST*) is one of the most well-known datasets in sentiment analysis [93]. In this scenario we use 6228 training and 692 validation samples about movie reviews from *SST-2*. Lastly, the Text REtrieval Conference Question Classification dataset, or *TREC* for short [94], contains 4906 labeled questions in the training set and another 546 for the validation one. It has 6 labels and the average length of each sentence is 10 with a vocabulary size of

[5]https://www.nltk.org/

[6]https://nlp.stanford.edu/projects/snli/

8700. In both datasets, SST-2 and *TREC*, we omit the classification purpose they originally had. Moreover, in the three corpus we employ all training samples to obtain the token embeddings in the GMVAE optimization since the set size is significantly smaller than in *SNLI*.

To speed up the loading of data, we utilize the extension *hdf5* for saving the hidden vectors. Moreover, this way we avoid memory issues when loading the datasets in the set ups since the files with the hidden states have significant sizes.

2.6.1. Noisy Regularized Sequence to Sequence

NoR-seq2seq

In this model, we follow the network structure from [14], omitting the attention mechanism for now. Consequently, we will use a LSTM as the RNN unit, a bidirectional encoder, a depth of two layers in the networks and a hidden size of 1024 for each of them.

The configuration of this scenario pursues a seq2seq pre-training of 120 epochs and a fine-tuning of the regularized decoder for only 20 epochs after training the GMVAE. In the GMVAE, after different experiments validating the hyperparameters, we finally chose 1500 for the hidden dimension, 100 for z, 20 for w and a K of 10 MoG in the prior. The depth in the networks is 5 layers and the deviation, σ, of the posterior normal distribution in the decoder 10^{-4}. We saved the hidden states (encoder output) of all the training sentences, and trained the GMVAE for 100 epochs.

One of the problems of this first model is caused by the limitations of our baseline. Seq2seq is not suitable for dealing with complex and realistic datasets, that is, long sentences and a wide dictionary, since they encode the semantic and syntactic information of a whole sentence in a single vector, i.e. the encoder output. Notwithstanding, we present in this section some examples where the effect of the regularization layer can be evaluated.

In Table 2.1 we show some test sentences reconstructed by our model compared with the baseline, which is the pre-trained seq2seq model without any GMVAE stochastic layer. In both cases, we reconstruct with the most likely word. By its own, a seq2seq model fulfills its task if the dataset is not very complex, so we have restricted the results to that premise (refer to Section 2.6 for dataset details). Our model achieves its goal when the sentences are short enough, finding words that fit the holes, while the baseline fails more often in the task, repeating the previous or following word into the masked place ([MASK] token) when it seems not to predict anything better (examples 2, 3 and 5). In addition, our method is able to change other words in the sentence, even if they were not masked, so the overall construction has more sense (example 2). However, in complex scenarios (example 5), both tend to fail, above all our approach, with ungrammatical sentences.

a woman standing in a dark doorway , waiting to be let into the building .
a woman standing in a dark small game waiting to be let into the building .
a woman standing in a dark blue jacket waiting to be let into the building .
a man in an orange hat starring at something .
a man in an hat hat starring at many .
a man in an orange shirt performs at night .
the red car is ahead of the two cars in the background .
the red car is is of the cars cars in the background .
the red car is is of the street cars in the background
five people wearing winter jackets and helmets stand in the snow , with snowmobiles in the background.
five girls , winter jackets and helmets stand in the snow , with flowers in the background .
five soccer , winter teenager and others stand in the snow with this river in the background .
a large bull targets a man , inches away , in a rodeo with his horns , while a rodeo clown runs ...
a bull bull targets a man , petting away , in a bottle with his other , while a rodeo clown tries ...
a young boy move a shoeshine opponent head , wearing a blue with the girl , with two boys ...

Table 2.1: Examples of sentences reconstructed by the regularized seq2seq. The first sentence is the original one, with the observed words underlined, i.e. **no underlying means a missing word**. The second is the output of the baseline seq2seq pre-trained. Finally, we show our method. The words in red correspond to mismatches with the original sentence.

NoR-seq2seq with attention

In this scenario, we keep the same configuration from [14], but including the global attention mechanism. Next, we show the results for both options describe in Section 2.4.2.

In the **option 1**, the C-GMVAE is trained with consecutive pairs of hidden states from the whole training set during 100 epochs. We finally used a hidden dimension of 1500, 150 for the latent space of z and 50 for w. We configured $K = 20$ classes in the MoG and a σ of 10^{-2} for the decoder posterior. The number of layers on each of the modeled distributions was 6. During the training we selected a learning rate of 10^{-5}, a dropout of 0.3 and a batch size of 64.

In the **option 2**, the GMVAE is trained for 150 epochs with the same configuration as before. It only changes the graph as described in Section 2.2.1.

In both options, for the pre-training of the seq2seq, 30 epochs were enough since the attention mechanism eases the convergence of the model. After the training of the C-GMVAE and the GMVAE respectively, we fine-tuned the seq2seq decoder with the inclusion of the suitable stochastic layer as mentioned in Section 2.4.2 during other 30 epochs.

Table 2.2 shows the results of the two configurations proposed. We use the same dataset as previously but as the model is more powerful due to attention, we are able to

increase the percentage of masked tokens to be inferred (see 2.6.1 for details of this **second strategy**) without damaging the overall performance of the seq2seq. Moreover, later we extend the results for a larger text corpora.

a man in an orange hat starring at something .
a man in a black hat starring at something .
a man in a hard hat starring at something .
a man in a black hat starring at something .
a boston terrier is running on lush green grass in front of a white fence .
a gray terrier dog running on the green grass in front of a blue shack .
a gray terrier dog running through tall green grass in front of a red ball
a black dog is running through the grass grass in front of a red flag .
a girl in karate uniform breaking a stick with a front kick .
a man in a uniform throws a stick to his his kick .
a boy in a uniform with a stick in a large kick .
a man in a uniform kicking a ball up to his opponent .
five people wearing winter jackets and helmets stand in the snow , with snowmobiles in the background .
two men in winter jackets and hats stand in a large space with structure in the background .
a group of winter day at a stand in a snowy area with trees in the background
two men wearing winter clothing and hats stand on the snow covered street with flags open .
a man in a vest is sitting in a chair and holding magazines .
a man in a vest is sitting on a rock and looking out .
a man in a vest is sitting on a sidewalk and playing music .
a man wearing a vest is sitting on a wall and smoking a cigarette .
a mother and her young son enjoying a beautiful day outside .
a mother and her daughter are enjoying a wedding day outside .
a mother and her child are enjoying a hot day outside .
a mother and her children are enjoying a hot day outside .

Table 2.2: Examples of sentences reconstructed by the regularized seq2seq with attention following the same format as Table 2.1: original, baseline, options 1 and 2.

The results in Table 2.2 show how both designs fit our goal, generating new sentences and computing substitutes to the masked tokens that fit the gaps. All of the sentences that are exposed belong to the testing dataset and have been selected randomly. As opposite as in the first scenario in Section 2.6.1 (NoR-seq2seq), the generation of sentences has improved due to the attention mechanism, so both the baseline and our method perform better the reconstruction of sentences as was expected. Moreover, the **option 2**, regularizing the context vectors, not only imputes the masked tokens but also some other tokens in the sentence so the complete structure makes sense. For example, in the second sentence the word 'terrier' is removed and 'dog' is changed of position. More interesting is the third one, where 'kick' and 'stick' are deleted but 'kicking' appears as a conjugation of 'kick'.

To understand the diversity of solutions achieved with our model, we can examine not only the most likely imputed word, but also the top five. We focus in **option 2** for simplicity.

For example, in the first sentence, the baseline best options for ‘orange’ correspond to colours, however our method also infers the word ‘cowboy’ in the top 5. In the longest sentence, the forth, we found that even if the final reconstruction was not completely correct (neither in the baseline), our method achieves more varied candidates. In particular, the word ‘snowmobiles’ has the more likely alternatives [‘structure’, ‘furniture’, ‘each’, ‘it’ and ‘reflections’] for the baseline while ours are [‘flags’, ‘trees’, ‘umbrellas’, ‘people’ and ‘something’], which is a more diverse set that absolutely fits the previous word ‘with’ in the sentence.

Our results demonstrate that our proposal performs at least as good as the baseline but in many times is capable to improve generalization in the imputation of missing words. Even more, it can be seen as a way of data augmentation in the sense that builds new sentences, acceptable and different from the baseline choices. One of the advantages that we will see in these models regarding Transformers is the flexibility of changing the sentence length since they generate words sequentially and stop it with the End of Sentence (EoS) token.

Next we show other results obtained as an extension.

Additional results

Table 2.3 presents additional results from the **option 2** in the seq2seq with attention model using the *SNLI* dataset. Once again, we prove the efficacy of our method, even if the dataset gets more complicated. In this table we present different samples of sentences reconstructed from the masked template, following the same philosophy of the results in Table 2.2. The fifth example exposes an extreme case where it is only observed the first word, ‘a’, and both the baseline and our method infer completely different sequences but good alternatives at the same time.

2.6.2. NoRBERT

Due to the costly process of training from scratch a Transformer, to implement NoRBERT, we make use of the pre-trained base model from BERT described by [15], using a MLM objective. This version of BERT is composed of 12 layers, a hidden size of 768 and 12 heads and we keep the parameters eased by the *Hugging face* library[7], which calls it *bert-base-uncased*. We keep the original configuration following the paper [15] except for the hyperparameters mentioned in the next sections. For RoBERTa and XLM-R models, following the same approach, we use the pretrained models called *roberta-base* and *xlm-roberta-base* respectively.

On each experiment, we train a GMVAE using the hidden vectors at some point of BERT structure obtained from training samples with the previous base models. Once the GMVAE has converged[8], we build a new architecture based on BERT with the integration

[7] https://huggingface.co/

[8] We consider the GMVAE has converged when the ELBO stabilizes during the training.

an old man with a package poses in front of an advertisement . an old man is standing with arms in front of an audience . an old man in a blue shirt in front of an audience .
a man playing an electric guitar on stage . a man playing an electric guitar on stage . a man plays an electric guitar and sings .
a blond-haired doctor and her african american assistant looking threw new medical manuals . a man is standing in an american assistant , using a medical apparatus . a man is looking at the american nurse to get a medical patient .
a young family enjoys feeling ocean waves lap at their feet . a young boy is feeling ocean and is on the beach . a young man in feeling ocean is surfing on a surfboard .
a man reads the paper in a bar with green lighting . a man is standing in front of a crowd of people . a man is sitting on a bench reading a book while sitting
three firefighter come out of subway station . three people come down a street corner . three people come out of a boat .
a person wearing a straw hat , standing outside working a steel apparatus with a pile of coconuts on the ground . a man wearing a straw hat , standing outside of a steel structure with a blue umbrella laying on the ground . a man wearing a straw hat , standing outside a large steel structure with a tree in front of the ground .

Table 2.3: Additional examples of sentences reconstructed by the regularized hidden states in the seq2seq with attention. Sentences order: original, baseline and reconstruction from our regularized option 2 of the GMVAE and the context vectors.

of a stochastic layer in the corresponding place of the hidden vectors. This new layer consists on the reconstruction of the hidden vectors through the generative network of the GMVAE. Finally, we fine-tune this new architecture, freezing all parameters below the stochastic layer in the computational graph.

Deep NoRBERT

First, we present the results of Deep NoRBERT, in which the GMVAE stochastic layer is placed in an intermediate BERT encoder layer, see Section 2.5.2 for more details. We show the results obtained in terms of accuracy and BLEU score for different locations of the GMVAE layer inside the BERT structure trained in the *SNLI* dataset.

The GMVAE layer is trained for 500 epochs with a learning rate of $5 \cdot 10^{-5}$. The GMVAE latent dimension z is set to 150, the w dimension to 50, and we consider a mixture of 20 Gaussians, dropout probability 0.3 and networks with a depth of 6 layers. Then, Deep NoRBERT is trained for 8 epochs freezing the parameters below the stochastic layer. The baseline BERT is also fine-tuned in the same dataset for 8 epochs so we can make a fair comparison in their performance in missing data imputation. We evaluate the percentage of tokens that are exactly the same as the source sentence in a 1-by-1 comparison. We test

two different scenarios, with masked tokens and with disrupted tokens, that is, instead of using the [MASK] token which indicates 'unknown', we place random choices from the vocabulary that damage the source sentence. We replicate the random words substituted on each experiment maintaining the same seed in the training. Regarding the masks, 40% of the sentences chosen at random have at least one [MASK] token, which always replaces a meaningful word (we avoid masks over stopping words).

Table 2.4 shows the **imputation accuracy** for different configurations, in which *l*-Deep NoRBERT means that we placed the GMVAE layer in the *l*-th transformer layer. For a better visualization, we highlight in bold every case that outperforms the baseline. Observe that the largest gains are obtained when the GMVAE layer is placed in the bottom of the network, outperforming BERT after fine-tuning. We remark that BERT is a state-of-the-art model for NLU that is pre-trained over a massive dataset and hence any improvement is not negligible, particularly when is achieved by placing a single regularization layer within. Despite some studies about BERT state that the last layers encode task-specific features [95], our results demonstrate that fine-tuning with the regularization of deep layers may improve the overall performance.

Model	Masked	Disrupted
BERT	97.13%	96.98%
1-Deep NoRBERT	**97.32**%	**97.11**%
2-Deep NoRBERT	**97.20**%	**97.07**%
3-Deep NoRBERT	**97.18**%	**97.1**%
9-Deep NoRBERT	96.87%	96.25%
11-Deep NoRBERT	96.05%	95.34%
12-Deep NoRBERT	95.89%	93.89%

Table 2.4: Accuracy of different models comparing the unmasked source sentence with the reconstruction. We evaluate a version that keeps the [MASK] tokens and other (disrupted) that substitutes them by random tokens from the vocabulary. In *l*-Deep NoRBERT, *l* refers to the transformer BERT layer in which the GMVAE is placed. The lower the deeper and the higher the closer to the classification top layer.

Table 2.5 presents the **BLEU** score obtained by Deep NoRBERT with different layer configurations. We explore different policies of generating missing tokens. 'Low' refers to the same mechanism as in Table 2.4 experiments. In the policies called 'Medium' and 'High' we do not exclude any token by its grammatical meaning (as it is done with stopwords before) and mask every word independently with probabilities of 0.4 and 0.6 respectively. Table 2.6 results, called **Masked BLEU**, differ from the previous ones in the n-grams taken for the metric computation. That is, we only consider n-grams that include a masked token. From both tables we draw similar conclusions: the best performance is obtained when the GMVAE layer is placed at the bottom of the network, right after the first transformer layer.

Model/Missing rate	Low	Medium	High
BERT	86.07	49.43	25.14
1-Deep NoRBERT	**86.90**	**49.91**	**25.53**
2-Deep NoRBERT	**86.65**	**49.75**	**25.26**
3-Deep NoRBERT	**86.53**	49.33	**25.45**
9-Deep NoRBERT	85.52	46.04	21.47
11-Deep NoRBERT	83.89	43.34	19.28
12-Deep NoRBERT	80.77	40.83	17.16

Table 2.5: BLEU score of different models comparing different missing rates.

Model/Missing rate	Low	Medium	High
BERT	3.73	21.3	15.34
1-Deep NoRBERT	**3.88**	**22.7**	**16.44**
2-Deep NoRBERT	**3.88**	**22.50**	**16.22**
3-Deep NoRBERT	**3.90**	**22.28**	**16.56**
9-Deep NoRBERT	**3.87**	19.78	13.34
11-Deep NoRBERT	3.65	18.21	11.64
12-Deep NoRBERT	3.01	16.31	9.61

Table 2.6: Masked BLEU score of different models comparing different missing rates.

Deep NoRBERT for other NLU tasks

In this section, we include an exhaustive study of 1-Deep NoRBERT as a solution to improve the BLEU score across different NLU tasks. In all cases, we check the validation score at several epochs during training. More precisely, in Table 2.7 we show results from *bert-base-uncased* (the baseline model used in previous experiments), *roberta-base* and *xlm-roberta-base*, which correspond to BERT, RoBERTa and XLM-R respectively as mentioned at the beginning of this section.

The way of completing the experiments is the following. First, we pretrained the baseline for 10 epochs. Second, we train the GMVAE with the embeddings from layer 1 (after the pretraining) in the suitable transformer model. Finally, we finetune NoRBERT for the number of epochs indicated minus 10. For example, in the first column of results, 40 epochs correspond to 10 epochs of pretraining plus 30 epochs of finetuning. In the baseline rows, the epochs are continuous without this partition. With this procedure, we make sure we compare the validation score at the same point in the timeline of training. It is important to mention that as we are interested in validation BLEU score and non the training one, as in other regularization mechanisms we deactivate the NoRBERT layer during evaluation.

With regards to the configuration of the GMVAE, we used similar hyperparameters as in previous sections. We validated the variance in a set from 10^{-5} to 1 in steps of one order

of magnitude. We kept the best configuration which is a deviation of 10^{-4}. We used a z dimension of 150, w dimension to 50, a mixture of $K = 20$, hidden dimension of 1500, 6 layers and 0.3 of dropout probability as we used in previous sections. The learning rate was set to $5 \cdot 10^{-5}$ and the number of epochs was set to 2000, 4000 and 5000 for datasets *Multi30k*, *SST-2* and *TREC* respectively. As the size of the dataset decreases, the training of the GMVAE becomes longer as we may sense since we treat each token as a independent input.

All improvements over the baseline are highlighted with bold writing in Table 2.7. In every case, our model overpasses the baseline, except for three cases with the *xlm-roberta-base* model. Therefore, it can be noticed that our model is more resilient to overfitting and this table, then, demonstrates the regularization properties of NoRBERT GMVAE layer. In both scenarios the validation score decreases with the number of epochs, but from the beginning our regularization improves the BLEU score in almost all situations.

Dataset	**Base**	**Model**	**Epochs**		
			40	**60**	**85**
Multi30k	bert-base-uncased	Baseline	0.834	0.841	0.831
		NoRBERT	**0.855**	**0.851**	**0.847**
	roberta-base	Baseline	0.865	0.859	0.855
		NoRBERT	**0.877**	**0.874**	**0.869**
	xlm-roberta-base	Baseline	0.893	0.885	0.886
		NoRBERT	0.892	**0.888**	0.878
SST2	bert-base-uncased	Baseline	0.834	0.82	0.804
		NoRBERT	**0.846**	**0.836**	**0.828**
	roberta-base	Baseline	0.857	0.85	0.83
		NoRBERT	**0.869**	**0.865**	**0.858**
	xlm-roberta-base	Baseline	0.874	0.871	0.867
		NoRBERT	**0.884**	0.867	**0.878**
TREC	bert-base-uncased	Baseline	0.854	0.828	0.815
		NoRBERT	**0.882**	**0.877**	**0.864**
	roberta-base	Baseline	0.883	0.871	0.845
		NoRBERT	**0.903**	**0.896**	**0.893**
	xlm-roberta-base	Baseline	0.881	0.849	0.836
		NoRBERT	**0.887**	**0.87**	**0.871**

Table 2.7: BLEU score in the validation set of several datasets and models comparing baseline with 1-Deep NoRBERT and varying the epochs during training.

Table 2.8 shows similar results, with the difference that we vary the rate of generating missing tokens during the MLM. Regarding the number of epochs, it has been fixed to 40. In this scenario the results are not so successful. Again, improvements over the baseline are in bold. However, we can conclude from this table that 1-Deep NoRBERT is not the best tool to deal with corpus that include a high rate of missing tokens. In other words,

Dataset	Base	Model	Missing rate				
			0.2	**0.3**	**0.4**	**0.5**	**0.6**
Multi30k	bert-base-uncased	Baseline	0.803	0.724	0.626	0.538	0.441
		NoRBERT	**0.811**	**0.725**	0.625	0.534	0.436
	roberta-base	Baseline	0.828	0.746	0.654	0.552	0.451
		NoRBERT	**0.831**	**0.747**	0.651	0.552	0.447
	xlm-roberta-base	Baseline	0,854	0.778	0.694	0.594	0.497
		NoRBERT	**0.859**	**0.784**	**0.699**	**0.596**	0.497
SST2	bert-base-uncased	Baseline	0.788	0.697	0.589	0.487	0.393
		NoRBERT	**0.8**	**0.705**	**0.596**	**0.49**	0.392
	roberta-base	Baseline	0.817	0.728	0.627	0.519	0.418
		NoRBERT	**0.827**	**0.736**	**0.638**	**0.526**	0.418
	xlm-roberta-base	Baseline	0,838	0.75	0.655	0.564	0.464
		NoRBERT	**0.841**	0.749	0.655	0.559	0.457
TREC	bert-base-uncased	Baseline	0.81	0.737	0.667	0.575	0.487
		NoRBERT	**0.845**	**0.764**	**0.685**	**0.59**	**0.493**
	roberta-base	Baseline	0.86	0.782	0.692	0.602	0.498
		NoRBERT	**0.87**	**0.788**	**0.696**	0.593	**0.502**
	xlm-roberta-base	Baseline	0.837	0.767	0.689	0.595	0.512
		NoRBERT	**0.854**	**0.779**	**0.694**	**0.598**	0.51

Table 2.8: BLEU score in the validation set of several datasets and models comparing baseline with 1-Deep NoRBERT and varying the missing tokens rate.

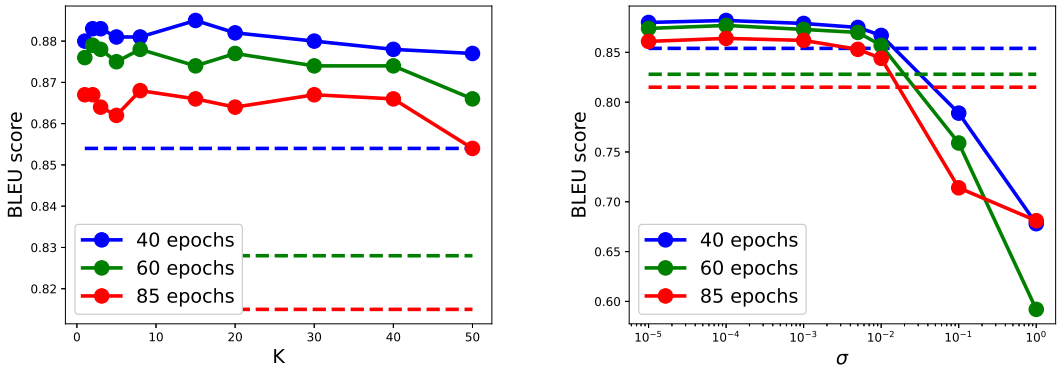
as we increase this rate, the BLEU score decreases and the differences between baseline and NoRBERT diminish, with a trend from the baseline to be better in high missing rates. In conclusion, for this scenario we encourage the use of Top NoRBERT instead of Deep NoRBERT as we show it to be more versatile after in this paper.

GMVAE validation

In this section, with 1-Deep NoRBERT, we validate in the *TREC* dataset two of the most important hyperparameters from the GMVAE, the variance of the posterior, σ^2, through the standard deviation, σ, and the number of components in the mixture of Gaussians, K.

On the one hand, in Figure 2.10a we appreciate the values of the BLEU score in the validation set for different number of epochs during training, as we did previously in this section. We compare each scenario with the baseline which is *bert-base-uncased* with no regularization layer included. As we excepted due to previous results, the score decreases as the number of epochs increases provoked by the overfitting. However, the variation of K does not affect to the results despite the fact that augmenting the number of components is related to more complexity in the model. We used 20 components in the experiments

because it is a value which maintains a trade-off between complexity and score.



(a) Study of K with $\sigma = 0.00001$.

(b) Study of σ with $K = 20$.

Figure 2.10: Validation of some hyperparameters in the GMVAE compared to the corresponding baseline scores (in dash linestyles).

On the other hand, in Figure 2.10b we show the effect of varying the deviation in the model. This hyperparameter has a more important role since it is related directly with the variance of the reconstructed embedding vectors. Therefore, the higher the deviation, σ, the higher the quantity of 'noise' that we inject in that input. If the variance is too small, the effect of the regularization layer is none. Nevertheless, if it is too big, we deviate the vectors too much from their hidden space, imputing tokens far away from what is expected in the sentences generation and therefore damaging the BLEU score. During our experiments, we chose a deviation of 10^{-4}, corresponding to a maximum in Figure 2.10b.

Top NoRBERT

The above results demonstrate that retraining BERT or any variant when we include a GMVAE layer within may bring imputation improvement when the layer is placed deep inside the BERT network. From this perspective, placing the GMVAE layer in the top of the network, as we do in Top NoRBERT, lacks a priori of any interest. Actually, when we freeze all the parameters from the encoder layers and fine-tune only the classification layer we achieve an imputation accuracy of 77.14% (Masked) and 75.53% (Disrupted) in the *SNLI* dataset, far below the Deep NoRBERT performance in Table 2.4. A closer look to the actual imputed words by Top NoRBERT in different sentences led us to conclude that the final GMVAE layer placed right below the classifier promotes topic diversity in the imputation task, which would explain the severe drop in accuracy w.r.t. Deep NoRBERT. This result may be consequence of the fact that upper layers in BERT learn specific features that affect the token choice while the deeper layers pick up general characteristics of language.

Therefore, in order to visualize the effect of the GMVAE at the top layer, Table 2.9 includes some test sentences reconstructed by Top NoRBERT in comparison with the

baseline BERT. Later we have included more examples with longer sentences (Table 2.10) as an extension. For the generation of the results we use again the *SNLI* dataset with the masking policy defined as 'Low'. The baseline corresponds to BERT model fine-tuned for half an epoch and a learning rate of $5 \cdot 10^{-5}$. The training of Top NoRBERT was fine-tuned with the same configuration. Regarding the GMVAE, we maintained all the previous parameters, except that we increased the learning rate to 10^{-4} and trained 200 epochs.

Source: This church choir sings to the masses as they sing joyous songs from the book at a church . **BERT:** this large choir looks to the camera as they sing joy about songs from the book at a church. **Top-NoRBERT:** a dancing band performs to the friends as they perform funcy bands from the book at a museum.
Source: A man reads the paper in a bar with green lighting . **BERT:** a man in the drink in a bar with green lighting. **Top-NoRBERT:** a man on the bike in a bar with green lights.
Source: During calf roping a cowboy calls off his horse . **BERT:** a the race a cowboy call off his back. **Top-NoRBERT:** during horse jumping a cowboy tries off his dog.
Source: A man in a black shirt is looking at a bike in a workshop . **BERT:** a man in a black shirt is looking at a woman in a conference. **Top-NoRBERT:** a man in a black shirt is looking at a sign in a shop.
Source: The man in the black wetsuit is walking out of the water . **BERT:** the man in the black wetsuit is coming out of the water. **Top-NoRBERT:** the man in the black swimsuit is jumping into of the water
Source: Five girls and two guys are crossing a overpass . **BERT:** Five girls and two guys are crossing a overpass . **Top-NoRBERT:** three girls and two guys are down a intersection side walk.

Table 2.9: Examples of sentences reconstructed by Top NoRBERT. The first sentence is the original one, with the observed words underlined, i.e. **no underlying means a missing word**. The second is the output of the baseline, BERT fine-tuned. Finally, we show our reconstruction. The words in red correspond to mismatches with the original sentence.

As it is shown in Table 2.9, the GMVAE stochastic layer at the top of BERT helps it to reconstruct sentences from a robust space, inducing the generation of more diverse sequences than the baseline. It is interesting how it changes some words maintaining the original structure as in the first example in Table 2.9. Moreover, these alterations maintain grammatical rules ('performs' and 'perform' are used according to the subject) and sometimes correspond to synonymous or analogous words (in this same example, the verb 'sing' is replaced by 'perform', the noun 'choir' by 'band', the object 'masses' by 'friends' and the place 'choir' by 'museum'). This diversity skill is not obtained by the baseline, so it is a characteristic uniquely from our methodology. In other cases, we get changes in words that are not masked so the overall sentence makes sense. The fifth example changes 'out' by 'into' as a consequence of infering 'jumping' from the masked word 'walking'. In the last example, NoRBERT changes 'crossing a overpass' by 'down a intersection sidewalk' as a semantically related structure that also corresponds the verb 'to

be'.

Enhancing diversity in text generation is a little explored area, as we do not even dispose of clear metrics to measure such an ability, in opposition to for instance image generation, in which researches typically rely on feature space metrics such as the FID to evaluate generation diversity [96]. We believe that the Top NoRBERT strategy to achieve such diversity may open future research lines on this topic.

Table 2.10 is an extension of Table 2.9 with results also from Top NoRBERT.

A man looking over a bicycle 's rear wheel in the maintenance garage with various tools visible in the background .
a man looking over a bicycle's back wheel in the maintenance garage with his tools visible in the background.
a man looking over a bicycle's rear wheel in a construction garage with wooden equipment is in the background.
A person dressed in a dress with flowers and a stuffed bee attached to it , is pushing a baby stroller down the street .
a person dressed in a suit with flowers and a stuffed animal attached to it, is pushing a baby stroller down the street.
a person dressed in a shirt with flowers and pink stuffed toy over to it, is riding a baby stroller down the street.
A blond-haired doctor and her African american assistant looking threw new medical manuals .
a blond - haired doctor and her african american doctor looking at new medical scrubs.
a blond - haired nurse and her african asian owner looking around new medical equipments.
3 young man in hoods standing in the middle of a quiet street facing the camera .
3 young man in hoods standing in the middle of a busy street facing the camera.
a young man in sunglassess standing in the front of a busy street holding the camera.

Table 2.10: Additional examples of sentences reconstructed by Top NoRBERT. The first sentence is the original one, with the observed words underlined. The second is the output of the baseline BERT fine-tuned. Finally, we show the reconstruction. The words in red correspond to mismatches with the original sentence.

2.7. Applications: Data augmentation

During this section, we want to emphasize a possible application of NoRBERT as a data augmentation tool. That is, due to the ability of our model to generate new sentences, we might focus in a NLP task and try to improve its performance with an augmented dataset. For this purpose, we have studied a classification problem in different datasets and we have compared a baseline model by Transformers, BERT with the original dataset and another one with augmented samples by our model.

The goal of NoRBERT in this task is the generation of new comments or sentences with some similarities to the base ones. As we are in a classification task, we do not want to change the label of the original sentences since it could damage the overall performance of the model but we want the new sentences to be different enough to add some information. We follow different steps for the training of our model:

- We train NoRBERT in a suitable dataset following steps in Section 2.5 (pretrain BERT - train GMVAE - finetune NoRBERT).

- We generate new sentences with NoRBERT, a desired masking rate and the dataset from the classification task, which can be the same or different from the previous step.
- We incorporate the new generated samples to the original dataset and train and evaluate the classification model.
- We compare the classification score with the same model trained in the original classification dataset without augmentation.

Regarding the baseline in the classification task, we will use *bert-base-uncased* from *Hugging face*, as we have done before in this chapter from pretraining NoRBERT. It is also the same model we use with the augmented samples. We will train it for an enough number of epochs and save the best model that is considered to be the one with the higher F1 score in the validation set. The training of NoRBERT constitutes the series of steps described in Section 2.5 and during these experiments we will vary the variance in the posterior distribution of the GMVAE, the number of epochs in the pretraining and finetuning, the datasets used on each stage and we will play with the quantity of samples used from the original dataset and with deep or top NoRBERT. The rest of hyper parameters stay as we studied and decided in previous experiments.

The F_1 score is a measure related to the classification accuracy and combines the precision and recall values. **Precision** answers the question 'how many selected items are relevant' so it is the total quantity of true positive results divided by the quantity of all positive classified results. **Recall**, on the other side, answers 'How many relevant items are selected?' and consists on the number of true positive results divided by all the results with positive true label, correctly or not classified samples. The next equation exposes its formula.

$$F_1 = \frac{2}{recall^{-1} + precision^{-1}} \tag{2.6}$$

In multi-class scenarios, we have used the extended version of F_1 score with the micro averaging strategy, that is, biased by the class frequency.

2.7.1. Model

As a variant, we introduce **Contextual NoRBERT** with the goal of incorporating more information in the generation of samples. As we presented previously, only Top NoRBERT was able to impute sentences from a more diversed topic space. However, with contextual information, Deep NoRBERT makes the work too. We will prove that both Contexual 1-Deep NoRBERT and vanilla Top NoRBERT are able to generate new samples that improve the classification accuracy on a secondary task. For convenience, we will use Deep NoRBERT when referring to 1-Deep NoRBERT in the rest of section.

Contextual NoRBERT differs from vanilla NoRBERT in the input and segment reconstruction but the architecture is the same. We select a contextual length and use the embedding of the words before and after the current word throughout that length. So, we concatenate the embedding of the contextual words of the current one as the input of the GMVAE but we do only focus in the reconstruction of the current word embedding. For example, when we use a length of 5 (contextual 5), we concatenate the 5 embeddings before and the 5 after to the current token and use all for the reconstruction of the segment in the input belonging to only the embedding of the current token.

2.7.2. Results

For this section, we have performed many experiments validating all parameters but we only show here the best cases and the configurations that worked. All results are computed from 3 to 5 times depending on the experiment. We use different seeds and present the average mean and standard deviation. Moreover, we always study the effects on the masking rate, since the more the masked tokens the more different the output sentence will be from the original one, but we take the risk that it has more grammatical errors or less sense. Furthermore, target information (the information that decides the sentence label) may be collected in one token and this token can completely change in the augmentation reconstruction. Therefore, some false positive/negatives could be created by mistake and damage the model in the training. The masking probability is then, undoubtedly, the most important parameter in this task.

Contextual Deep NoRBERT augmentation in SST-2

We first expose the results using Contextual Deep NoRBERT and SST-2 dataset on its binary classification task, where its goal is to classify movie reviews as positive or negative comments. SST-2 is composed of 6228 training samples, 1821 validation samples and 692 test samples. We pretrain and finetune NoRBERT during 30 epochs in 50000 sentences from OSCAR dataset because of its large size and complex sentences so our model can benefit from language structures learnt from other databases. Regarding the configuration of the GMVAE, we set the same as in previous sections with a variance in the posterior of 10^{-4}. The generation of sentences is performed in separated runs per class, that is, we take the samples with positive labels and generate new positive ones, and the same with the negative sentences.

With results from Table 2.11 we can see that the data augmentation by our model is a better tool in scenarios where the number of samples is reduced since the F_1 score is improved in all cases when using 1% of the samples and contextual 5, and almost all of them in contextual 10. However, the deviation of these results is too high if we compare it with the 100% samples scenario where the score is more stabilized. In that case, the improvements in F_1 score are no so good although it is possible to obtain better score

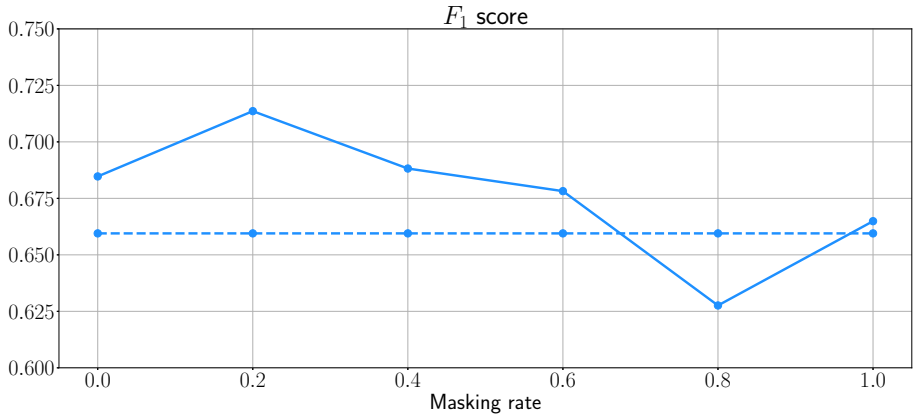
in some configurations. Anyways, we can not confirm any statement about the masking probability from that table.

		Contextual 5				**Contextual 10**			
		mean		deviation		mean		deviation	
		test	val	test	val	test	val	test	val
	Baseline	0.6595	0.6798	0.0261	0.0294				
	0	**0.6641**	**0.6886**	0.0344	0.029	0.6379	0.6634	0.1037	0.1004
	0.2	**0.6683**	**0.6921**	0.0608	0.0557	**0.662**	**0.6848**	0.009	0.0113
1%	0.4	**0.6615**	**0.6874**	0.0205	0.0207	**0.6821**	**0.7046**	0.0392	0.0313
	0.6	**0.6692**	**0.6865**	0.0213	0.0091	0.6556	0.6629	0.0641	0.0808
	0.8	**0.6646**	**0.6834**	0.0661	0.0695	**0.6611**	0.676	0.0486	0.0616
	1	**0.6691**	**0.6851**	0.0292	0.0216	**0.6746**	**0.6848**	0,0058	0.0191
		test	val	test	val	test	val	test	val
	Baseline	0.9161	0.918	0.0056	0.0085				
	0	0.9116	**0.9216**	0.0028	0.0076	0.9081	**0.9198**	0.0082	0.0021
	0.2	**0.9213**	0.9175	0.0026	0.0069	**0.9161**	0.9142	0.0079	0.0038
100%	0.4	0.9149	0.9133	0.0039	0.007	**0.9177**	0.9162	0.003	0.0032
	0.6	0.9121	0.9167	0.0072	0.0034	**0.9190**	**0.92**	0.0054	0.005
	0.8	0.9113	**0.9186**	0.0031	0.0038	**0.9182**	**0.919**	0.0058	0.0013
	1	**0.9186**	**0.919**	0.0066	0.0047	0.9154	0.9168	0.0053	0.0056

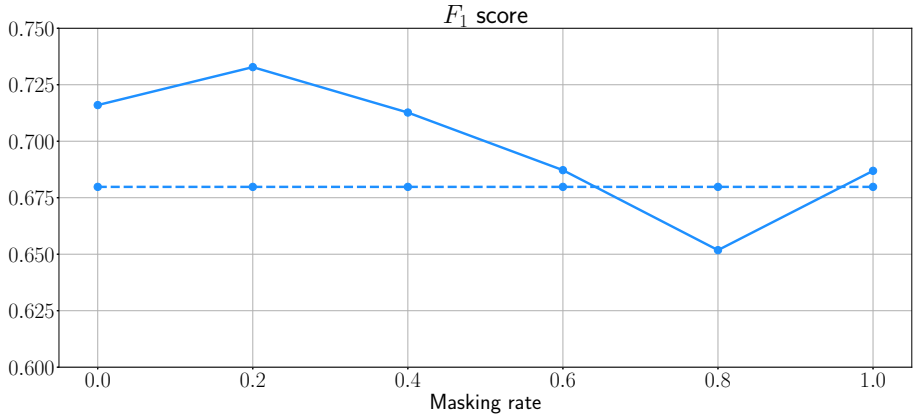
Table 2.11: F1 score using 1% and 100% of samples in the classification of SST-2 data. It is used Contextual Deep NoRBERT with size 5 and 10 in the embeddings window. We show the mean and the deviation of 3 different runs for each experiment. Moreover, we vary the masking rate from 0 to 1 in steps of 0.2.

Then, we try to improve the training a little bit by doing a second finetuning in NoRBERT with the samples from the classification task, in this case, SST-2, during other 30 epochs. So, if we are using only 1% of the total samples, we will use also that same set in this second finetuning so we do not fool the model. These results are presented in Figure 2.11 and they show a better prove of the variation in the masking rate. With that figure information we can state that larger masking rate degrade the system due to the augmentation with bad samples, that is, sentences without grammatical structure or very different from the original ones. If we use a masking probability of 1, we are masking all words in the source, so NoRBERT has no information to generate a new sentence and the output tends to appear as a sequence of random tokens. On the other extreme, if we mask no words, masking probability of 0, even though NoRBERT is able to change some tokens due to the GMVAE layer, it is less likely and the generated sentence might be equal to the input. Finally, we can confirm that rates around 0.2 seem to be the perfect trade-off since it is low enough to maintain original sentence meaning but high enough to support different connotations.

Due to the problematic with possible label changes in the data augmentation, we



(a) Test set



(b) Validation set

Figure 2.11: Averaged F_1 score versus masking rate with 1% of samples and data augmentation by Contextual 5 Deep NoRBERT compared with baseline (no data augmentation at all). Dashed lines means baseline.

made a quick experiment finetuning NoRBERT with SST-2 comments with their label concatenated. With this procedure, we let the model change the label token as well as the rest of words in the review. In Table 2.12 we expose several comments in the original form and the reconstruction from Deep NoRBERT. As we train the regularized transformer with the label information concatenated as the first token, it learns to generate a new label in that position with the two possible values (positive or negative). We see from the examples in that table that when the label changes, some tokens in the comment also change to follow the connotation transformation. However, there are some cases where the label generated is not correct (not any of the two choices) so we have to discard the sentence. That may be the reason of the final scores, that do not improve the ones showed in Table 2.11 so we do not present them. Then, we keep for future research an improvement of data augmentation task with the choice of change label with a deeper study.

Label	Comment (original - reconstruction)
negative	hard as this may be to believe, here on earth, a surprisingly similar teen drama, was a better film.
positive	you of this may hard to believe, here on earth, a surprisingly good teen drama, is a great film.
negative	both deserve better.
positive	it even better.
positive	elling builds gradually until you feel fully embraced by this gentle comedy.
negative	elling builds suspense until you feel like it a a romantic comedy.
positive	it's definitely an improvement on the first blade, since it doesn't take itself so deadly seriously.
negative	it's be an improvement on the first try, since it doesn't take itself so deadly seriously.

Table 2.12: Examples of label change in the reconstruction by Deep NoRBERT.

Comparing Top and Deep NoRBERT augmentation in SST-2

In the study with SST-2 dataset we also include a set of experiments where we compare Top and Contextual 5 Deep NoRBERT data augmentation in a scenario with different number of samples. The only difference regarding the previous configuration of the hyperparameters is the posterior variance of 10^{-2} of Top NoRBERT regarding 10^{-4} of Deep NoRBERT and 4000 training epochs instead of 2000 to ensure the GMVAE convergence. In this setup we consider a class-balanced problem with few samples per class. In Figure 2.12 we present the F_1 values when the number of samples per class goes from 30 to 90, or what is the same, from 60 to 190 samples in total. We do not show cases with less samples cause the results appear with very high variance and no conclusions are obtained. That is due to the lack of advantages in finetuning such a big model with so few samples and also in the randomness included by the GMVAE when generating new samples. Anyways, we may state a higher variance in Top NoRBERT results because of the more varied sentences generated. In both models from Figures 2.12a and 2.12b we see the tendency of decreased score with higher masking rate, what we stated in the experiments before. Moreover, we can also say that results with Contextual Deep NoRBERT are better in terms of consistency and classification score. Nevertheless, the choice of the model depends on the scenario, the kind of dataset set and the number of samples we are dealing with. When we have a bunch of samples, finetuning all layers above GMVAE regularization in Deep NoRBERT might be problematic and far from adding the generalization and diverse space we want to achieve.

Augmentation in other datasets

In this section we include as a final result the study of the augmentation in the classification of a different dataset with different number of classes. For this scenario we take all samples from dataset *AG news*. It is a collection of more than one million of news articles from more than 2,000 sources[9], labeled with four different classes regarding the category of

[9]http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

the document (world, sports, business and Sci/tech). It is composed of 7,600 samples for test and 120,000 for training, that we split in 96,000 and 24,000 samples for train and validation respectively.

In Table 2.13, we present the values for the F_1 score in the same models as before with Top and Deep NoRBERT. We keep the same configuration. We only evaluate the results for the masking rate that we have analysed that works the best, from 0.2 to 0.4. As we see in the table, in all scenarios, we overpass the baseline score, with the best policy the masking rate of 0.3. We achieve an improvement in the test set of 0.42 points and 3.29 points in validation. We run again three times the experiments and average them.

		test	**val**
Baseline		0.9425	0.9454
Top NoRBERT	0.2	0.9428	**0.9883**
	0.3	**0.9454**	0.9735
	0.4	0.945	0.9669
Contextual Deep NoRBERT	0.2	0.9466	**0.9783**
	0.3	**0.9467**	0.9732
	0.4	0.945	0.9688

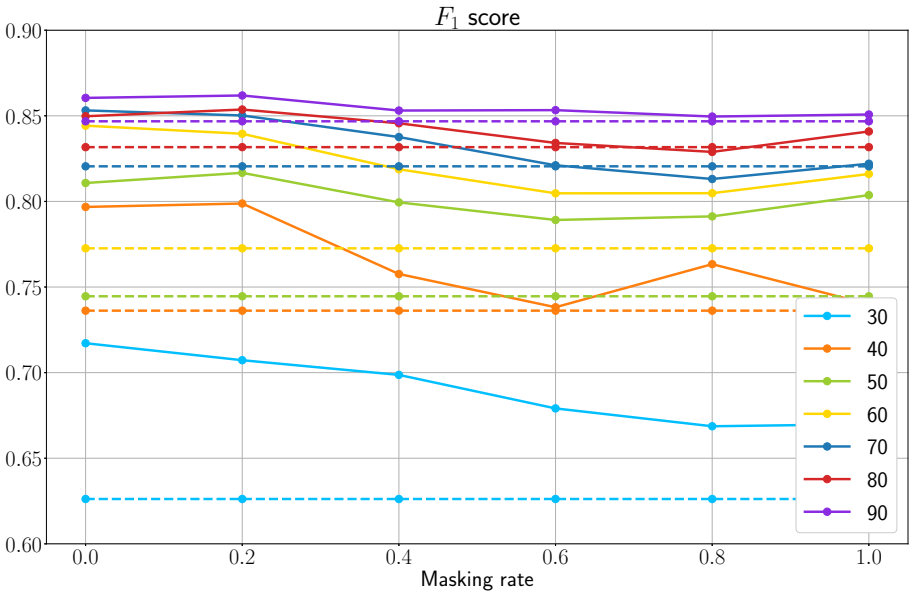
Table 2.13: F_1 score in AG's news dataset by Top and Deep NoRBERT. We have marked in bold the best cases in both models, Top and Deep NoRBERT.

2.8. Conclusions and future work

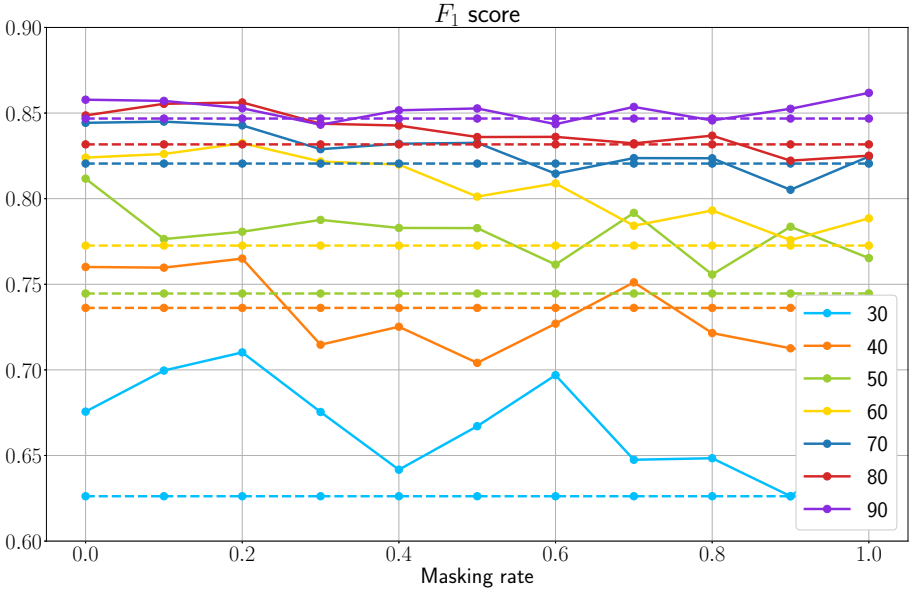
In this chapter we have proved the successful effect of adding a stochastic GMVAE layer in LMs through NoR-seq2seq and NoRBERT. On the one hand, despite its limitations with long sentences, regularized seq2seq is able to predict assorted structures upon an extend. Then, we enforced the same idea applying attention and exploring other scenarios that incorporate the regularization at different points of the baseline. In this work, we successfully reconstruct a varied set of topics from the masked source sentences and demonstrate the efficacy of the stochastic layer in finding synonymous or analogous fragments that fit in the gaps. On the other hand, in NoRBERT, we study the different advantages regarding the layer where it is applied. While Top NoRBERT also successes with an increment of diversity as well as an easier way of adaptability to new contexts, Deep NoRBERT responds better in terms of accuracy and BLEU score. In the former case, we propose a novel methodology to generate new structures of text with diverse topics that fit the gaps thanks to the inclusion of controlled noise through a DGM. As an extension, we prove Deep NoRBERT to improve the BLEU score in additional datasets and three different Transformer models. As a way of reinforcing our idea, we prove the GMVAE effect regularizing a well-studied scenario with FMNIST images. Finally, we include a secondary application based on the reconstruction of sentences to use them in

data augmentation. We explore the advantages of this mechanism in a classification task, obtaining better F_1 score when using NoRBERT augmented samples regarding not using augmentation at all. We show improvements in two datasets with different number of classes and percentage of samples. This application is also used as a measure to prove the effectiveness of our model.

For now, there is no metric to evaluate robust and varied solutions in word imputation problems (as NoRseq2seq or Top-NoRBERT achieve), since traditional evaluations as BLEU [62] or ROUGE [97] are based in the reconstruction of the original sentence. There is no perfect evaluation metrics testing the text generation because it is difficult to resume all the semantic and syntactic properties that language needs to fulfil [98]. Therefore, we let for future work the exploration of metrics or lost functions that allows the LM to generate sentence embeddings with more diversity based on the context. Furthermore, as another tool of evaluation, we also consider a future line of research the effects of NoRBERT in specific tasks as sentiment analysis or machine translation in addition to the classification task presented at the end of the chapter.



(a) Contextual Deep NoRBERT



(b) Top NoRBERT

Figure 2.12: Averaged F_1 score versus masking probability and different number of samples per class (from 30 to 90) of augmented model compared with baseline (no data augmentation at all). Dashed lines means baseline results (not affected by the masking rate).

3. PSYBERT, A TRANSFORMER APPLIED IN PSYCHIATRY

This chapter comprises the junction of bidirectional encoder representations from Transformers (BERT) within the NLP sector (studied in Chapter 2) and the line of health-related applications that we highlight through our work. Therefore, it will constitute the first medical objective we present but not the only one we analyse during the thesis. More precisely, in this chapter, we will use the EHR (Electronic Health Record) in order to **predict undetected or missing diagnoses** from patients with any mental disorder. For those tasks we will face the detection of errors in a labeled database, the filling of lost information or even scenarios with comorbidity (the presence of several diagnoses simultaneously). Then, we need a robust model, capable of dealing with all these factors in a heterogeneous data space.

In this context, we propose PsyBERT, a transformer-based architecture that combines embeddings from different heterogeneous EHR fields to study the diagnosis of mental disorders. We are inspired by BEHRT [19] with the novelty of dealing with heterogeneous data, including continuous and categorical data and most important, medical natural language.

We divide the chapter in four sections. First, we make an introduction about health-related models and the problematic in the field of psychiatry and diagnosis. Then, we detail the data we are working with and the preprocessing we apply to it. Following, we present our model, PsyBERT, and its novelties beyond BEHRT. Finally, we address two relevant problems where PsyBERT has demonstrated clinically relevant results, missing diagnosis imputation (Section 3.4) and delusional case detection (Section 3.5).

3.1. Introduction

In the last decades, artificial intelligence has been used in psychiatry for several purposes such as detecting patients at risk of suicide [99], [100], predicting psychotic disorders [101] or psychotic relapse [102]. In particular, NLP is showing promising results in the sector [103] but some limitations such as the difficulty to obtain high-quality annotated databases or lack of studies in non-English scenarios still need to be solved. We propose in this work the analysis of EHRs for the diagnosis of mental health patients. However, we are limited by the data quality in these databases [104], [105] with usual incorrect diagnosis codification and the huge quantity of missing information [106], [107]. Hence the medical sector is the perfect candidate to exploit NLP tools thanks to the huge quantity of unstructured text is produced daily in hospitals without being explored yet [108].

Some of the most relevant works in the literature about NLP in psychiatry and neuroscience comprises open source tools for information extraction [109]–[111] or pipelines

for classification from clinicians' notes [112]–[114] or text produced by patients [115]–[117]. In other scenarios, the resulting knowledge can be harnessed to address multiple long-standing problems in psychiatry such as diagnostic instability [118], [119].

Among these models, Deep learning usually has the problematic of needing also big annotated datasets. Nevertheless, with Transformers, this double issue is solved since their pre-training stage allows an unlabeled general dataset as medical records and the finetuning can be focused in a specific task with fewer samples. Due to the similarities between text and EHRs regarding the sequential nature of data and the large size of the vocabularies, more and more models which combine EHRs and Transformers are appearing in the literature. Med-BERT [120] is a transformer-based model pre-trained over 20 billion of patient's EHR in order to predict both heart failure in diabetes patients and pancreatic cancer. It modifies the embedding construction of BERT combining code, visit and serialized embeddings and removing special tokens such as the classification one, [CLS]. Then it summarizes the outputs with a feed-forward layer (FFL) before applying the specific-task one. G-BERT [121] combines BERT with Graph Neural Networks (GNN). They integrate a GNN representation of the diagnoses codes into BERT and pre-train using patients with a single visit data. Later, they finetune in longer sequential registers to solve the medication recommendation task. Finally, SARD (Self Attention with Reverse Distillation) [122], applied to different clinical prediction problems, is inspired by BEHRT with the novelty of reverse distillation as the pre-training procedure.

BEHRT [19] was first introduced in 2020 by Yikuan Li et al. as a disease predictor through a modified BERT model and the sequential diagnoses and age data from the EHR. Then, other variants have been developed as Hi-BEHRT [123] which appeared as an improvement with the inclusion of more clinical information (medications, measurements...) and a hierarchical structure capable of dealing longer sequences, or Targeted-BEHRT [124] with the combination of static and dynamic EHR data among others contributions.

In this chapter we have also designed a model based on BEHRT that we have called PsyBERT and, when applied to electronic health records from psychiatric consults, enables the identification of patients with any general mental pathology or, more specifically, a delusional disorder. Take into account the importance of delusional case detection due to the misdiagnosed patients in the data. As we will see in the following sections, these tasks are achieved on the one hand with a MLM policy pre-training and, in the other, through a classification layer on top of the model that predicts the likelihood of having simultaneously each diagnosis in the last visit of the patient.

Currently, we are writing into two different papers both works presented in this chapter.

3.2. Data description

The samples we will study during this chapter come from the whole EHR (with 50 fields of information) of mental disorder patients from the Hospital *Fundación Jímenez Díaz*

in Madrid. This database is made up of 315,608 full registers and 420,232 registers with missing diagnoses, all from 46,238 unique patients. Each patient has a mean of 6.82 registers, and each register a mean of 1.24 diagnoses. With these numbers we would like to emphasize the proportion of patients with missing diagnoses regarding the total which is 57.1%, showing the problematic in the lost of information in today's EHRs. That is the main reason why we want to focus this chapter in solving this issue with different perspectives.

As a representative example of the dataset distribution, in Figure 3.1 we show the most common diagnoses found in the EHR from the patients. Appendix A includes more details about the *International Statistical Classification of Diseases and Related Health Problems*, ICD-10 codification [125]. The more frequent cases belong to dysthymia and personality disorder.

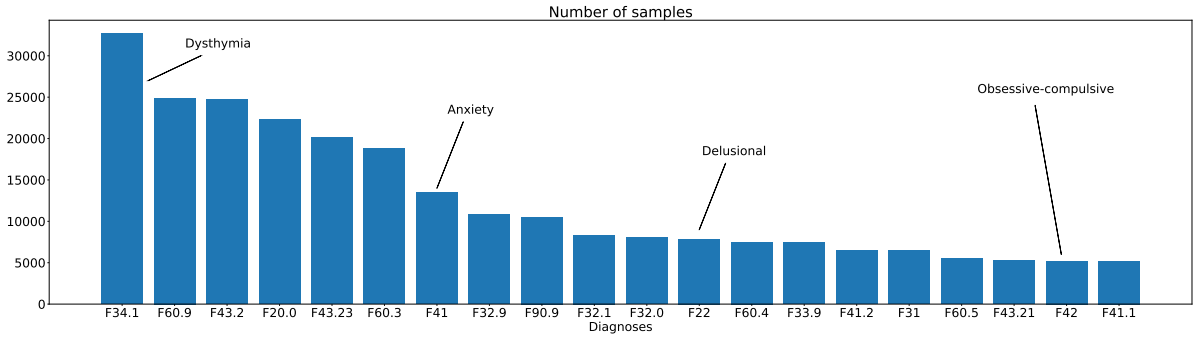


Figure 3.1: Most common diagnoses in the database regarding ICD-10 codification. The predominant ones, F34.1 and F60.9, belongs to persistent depressive disorder (dysthymia) and personality disorder respectively.

We consider each data point a sequence belonging to a patient, meaning that it is composed of several registers where each register may have one or more than one diagnosis. Hence, we face with two common situations in this kind of data, temporal sequences and comorbidity. More precisely, we will use the following fields from each register on each patient:

- **Date of the register**. It is used for sorting the database (sequence of visits) chronologically.
- **Patient's age**: We compute this feature subtracting the birth date from the register date. It is an integer indicating the age in a quantity of total months.
- **Sex**. Coded by a integer meaning 'male' (1), 'female' (2), 'transsexual' (3), 'transsexual male to female' (4), 'transsexual female to male' (5) and 'not sure' (6)
- **Diagnoses**: It is a list with a maximum of 8 diagnoses coded regarding the ICD-10 codification.

- **Discharge treatment**: It is a free text field. It refers to the treatment with which the patient discharges from the hospital after an admission.
- **Psychiatric history**: It is also a free text field. It collects the previous contact with Mental Health that the patient has had, previous diagnosis he or she has received and date, treatments taken in the past,... It usually has information about history with toxic consumption (if there is) and data related to social situation of the patient (marital status, people he or she lives with, job,...)

Field description	Percentage of samples
Reason for consultation	62.26%
Birth date	100%
Sex	100%
Register date	100%
Psychiatric history	88.69%
General history	3.36%
Degree	43.12%
Recent suicide planning	59.02%
Number of previous suicide attempts	59.8%
Degree of medical harm as a result of the current attempt	5.55%
Recent suicidal ideation	58.89%
Family history of suicide attempts	55.9%
ICD-10 coded diagnosis from axe I	14.42%
Diagnosis from axe I	77.68%
ICD-10 coded diagnosis from axe II	1.72%
Diagnosis from axe II	7.64%
ICD-10 coded diagnosis from axe III	0.84%
Diagnosis from axe III	6.26%
Diagnosis from axe IV	1.44%
Discharge treatment	88.48%
Psychopathological exploration	16.77%

Table 3.1: Percentage of samples and most relevant fields in the EHR. In this table there are mixed heterogeneous data such as categorical values (sex or coded diagnoses) and text (psychiatric history).

In Table 3.1 we show the most relevant fields in the EHR with the percentage of samples stored. It is an example of the problematic in today's EHR, since the missing data rate is significantly high in a great proportion of the fields. The most important issue is regarding the coded diagnoses. More specifically, diagnoses from axe I, the most important and representative one in the mental disorders we are studying, only have a 14.42% of coded diagnoses and a total of 77.68% diagnoses described by a free text field. The main issue of this scenario is the uncommon nomenclature to define a specific disease, finding

several definitions to name it. One of the goals in this work is to unify all these definitions and complete the missing diagnoses presented in the data base.

In relation to the definition of the different axes, the number I encompasses clinical psychiatric syndromes (e.g. schizophrenia, bipolar disorder, delusional disorder...). Axis II refers to developmental and personality disorders. Axis III encompasses physical illnesses. Axis IV is related to psychosocial and environmental problems. Finally, axis V, not shown in the table, is a scale that grades the functioning and activity of the person. Axes I, II and III include coded diagnoses, axis IV is a drop-down menu with a list of problems that can be selected (e.g housing problems, problems related to the environment...) and axis V is a scale that grades the patient's functioning and activity from 0 to 100 in 10-point intervals (e.g. 91-100 Satisfactory activity in a wide range of activities, valued by others because of its many positive qualities. No symptoms).

We also appreciate in Table 3.1 the main reason to select the free text fields from treatments and psychiatric history since they have the highest rates of present information with 88.48% and 88.69% respectively.

3.2.1. Preprocessing and dictionaries construction

Each sample in the dataset construction represents a patient and the information within each patient is sorted in a chronologically way. The different values we can find within the data points refers to age, sex, diagnoses, treatments and psychiatric history from the different visits along their clinical records. We add more or less features regarding the model we use, as we describe in Section 3.3.

The problem solving from the model we will use, PsyBERT, consists on two stages as we usually see in Transformers models and we dealed in Chapter 2. In the first stage we apply a MLM training for the whole dataset, all patients with their visits. In the second one, we finetune the model for a specific task, which, in this scenario, is the prediction of the diagnoses from the last register. Then we create these two dataframes, one with the whole sequence of registers per patient and the other one with all registers except the last one, separated as the label to be predicted from a supervised perspective.

In PsyBERT, the model we use and that is described in Section 3.3, each feature is added through an embedding layer so we create dictionaries for sex, ages and diagnoses. We only get two different sex values in the samples from this study, male and female, coded with 0 or 1. The maximum age we consider is 110 years in steps of one month. So, we convert date-type values in floats measuring the quantity of months, and then pass them to integers inside a categorical interval. Finally, the diagnoses are coded regarding the ICD-10 notation (See Appendix A) with a total of 768 different items. Note that from a supervised nomenclature, we will solve a classification problem with 768 different classes.

Regarding the text fields, *discharge treatment* and *psychiatric history*, we apply a standard preprocessing from language models that consists on the following steps with the

training of a tokenizer:

1. Lower case.
2. Remove punctuation.
3. Remove accents.
4. Remove stopwords.
5. Separate digits individually.
6. Apply a word piece tokenizer with a vocabulary size of 50,000 tokens.

The creation of the text dictionary is performed through the last step. We train a word piece tokenizer after the preprocessing and then save it so we can load it to code the input data in our PsyBERT. All these steps, with the inclusion of an explicit example, are shown in Figure 3.2.

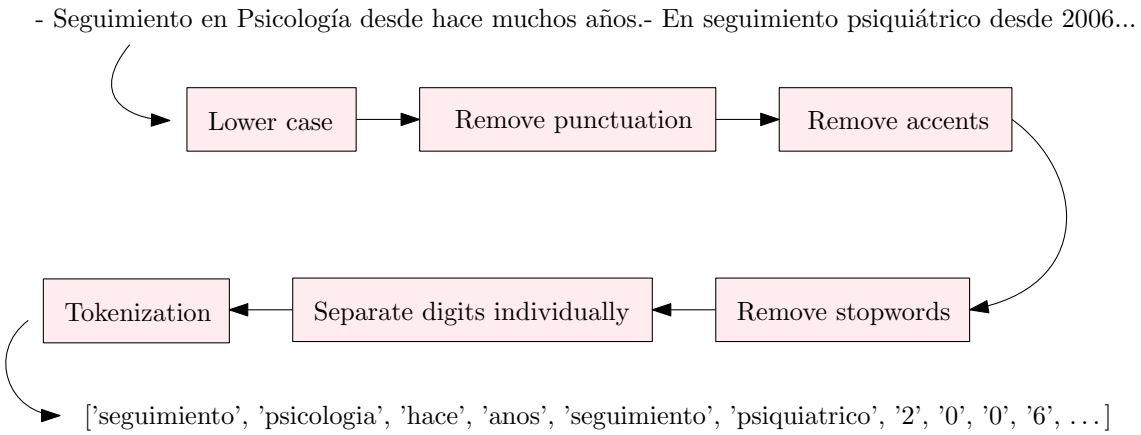


Figure 3.2: Text preprocessing steps for *discharge treatment* and *psychiatric history* fields in the EHR. We include an example from *psychiatric history* text.

The train-test split proportion we apply is a ratio of 80-20%, with 36,990 and 9,248 patients respectively.

3.3. Model description

3.3.1. BEHRT and its modifications

BEHRT[19] is an architecture based on BERT [15] adapted to work with the EHR. As we described slightly BERT model in Chapter 2, we will focus on the main differences with respect to it and, of course, in our contributions.

BEHRT was developed to predict the probability of 301 different conditions in future patients' visits. Its best characteristic is the ability to modify its structure to combine

heterogeneous concepts and the chance to provide an easy explainable output so important in health. The original model is built in a way that uses information from diagnosis and age to make the forecast. As an improvement, we integrate three additional heterogeneous features, including information in text format, and perform a prediction over more than the double of disorders from the original work.

Despite the fact that BEHRT's original goal is the prediction of diagnoses over three different time scenarios, the next visit, the next 6 months and the next 12 months, we completely change these tasks. We only take advantage of the pre-training step and then we make the prediction of the last visit to solve the problems we discuss in detail in Sections 3.4 (missing diagnoses) and 3.5 (detecting delusional patients).

The principal variation with regards to BERT structure falls on the embedding vector construction before applying the bidirectional encoder. However, the layers distribution works the same way with the multi-head attention (remember Figure 2.3) and the specific-task block. The latter is followed by a *softmax* function that gives the model the multi-label classification property and the ability to compute the likelihood of every disorder simultaneously. This may be the most important block in the structure since it gives the model the comorbidity properties as well as an output in terms of probability.

Even though Transformers appeared in the first instance as language models, many authors had already turned around this idea (see Section 3.1) and treated the EHR as a document, where each diagnosis is a word, each visit or register is a sentence, and the entire medical history of a patient is a document. The point of this process is the direct application of the MLM policy to the diagnosis in a pre-training stage and the consequently chance to perform a finetune with a specific-task layer on the top of the architecture as BERT or any other Transformer does.

3.3.2. PsyBERT and its embedding layer

The embedding layer is the principal block in PsyBERT architecture and it is also the one that we have modified the most with respect to BEHRT. It originally consists on a combination of four embeddings that we extend until six and has the ability to learn the whole evolution of a patient in a single embedding through a summation. The fixed embedding that we maintain are the disease, age, position and visit segment, which are the first, third, fourth and fifth embeddings in Figure 3.3. The other three are PsyBERT contributions and, more specifically, the procedure to integrate the text embedding is completely novel.

Diseases codes (DIAGS in Figure 3.3) are the main information in the model as they will also define the possible values in the output of the network (expect special tokens which are removed from the classification layer dictionary). **Age** represents a key concept since encodes two kinds of information, not only the epidemiological notion of when the event occurred but also the time between events as a sequential indicator. For example, in

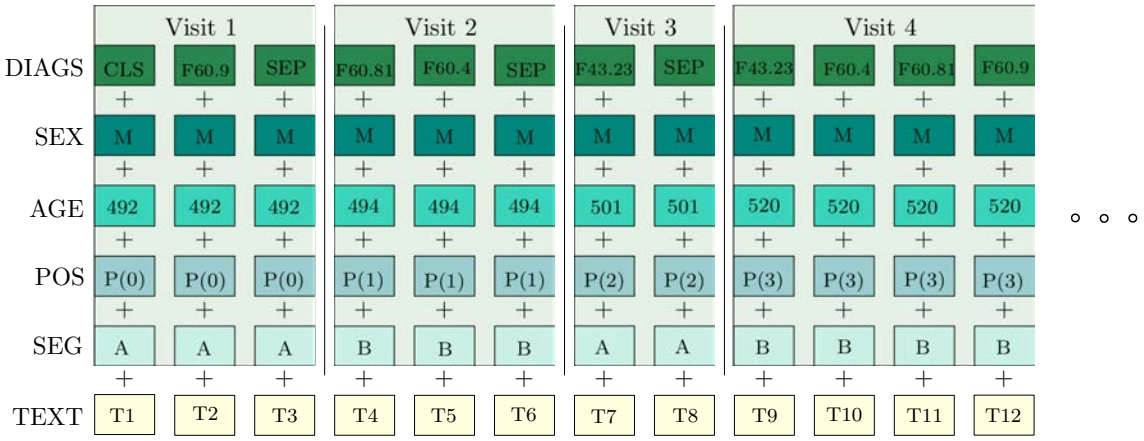


Figure 3.3: Diagram of PsyBERT embeddings. The final output is the summation of all the embeddings and it is the input to the bidirectional encoder. The dots in the right side of the image refer to the possible additional visits that the patient may have. Take into account that the length of the text sequence can be higher that the rest of visit-related features. In any case, padding tokens [PAD] are added until the block size previously defined in the hyperparameters configuration.

Figure 3.3 we have 492 and 494 months in the respective age values in visits 1 and 2, what defines an age of 41 years old in the patient and a distance of two months between visits. **Position** encoding (POS in Figure 3.3) has the objective of numerating the registers/visits in the EHR evolution of a patient and follows a sinusoinal function similarly to BERT. Finally, the visit **segment** can have two different values and mark the change of a visit to the next one alternating its value. In PsyBERT, we include the embeddings from the **sex**, where we find the possible choices 'male' or 'female'. Then, the only embedding that differs its value intra-visit is the diagnosis one, since all the others maintain it. That is, if we come back to the language analogy and take a visit as a sentence, it may be composed by several diagnoses due to comorbidity and each one will be consider a different word or token with a different embedding vector. However, the embeddings of the other fields will maintain their values per all corresponding diagnosis tokens in the visit. See for example visit 1 in Figure 3.3, with repetitions of 'M', '492', 'P(0)' and 'A' for sex, age, position and segment respectively. However, the diagnoses follow the regular concatenation rule in NLP with 'CLS' and 'SEP' tokens. In addition to this, we include a free **text** field that can be either 'discharge treatment' or 'psychiatric history' (see Table 3.1) as extra information for the model. We do not add both text embeddings in the same model but only one of them. This field is coded after concatenating the text sequences from all the visits in the clinical history of a patient. The tokenization is done through a pre-trained tokenizer in the preprocessed source of text as explain in Section 3.2.1. Hence the coded text (T1, T2, T3, T4... in Figure 3.3) is not associated to a unique visit but to the whole patient sample. Due to this long field, block size in our model must be increased to at least 512 to be considered a good parameter capable of collecting longer sequences of text.

3.3.3. Training methodology and configuration

PsyBERT as a Transformer model is divided into two training stages. The first one is the MLM that we will use as a tool to predict the missing diagnoses in the database in Section 3.4. This procedure makes PsyBERT learn the contextual representation of each diagnosis so we can then use it as a disease predictor when this field is missing. The second one is the specific-task step, where we include a multi-label classification layer to the previous pre-trained model to finetune it and predict the diagnoses of the last visit.

Regarding the parameters configuration we use during this work, they consist on a learning rate of $3 \cdot 10^{-5}$, a warm-up proportion of 0.1, a weight decay of 0.01 and a dropout rate of 0.1. As in the original work, we use a combination of 6 multi-head attention layers and 12 attention heads, what seems to work fine. In addition, the size of the intermediate layer in the encoder is 512 and the non-linear activation function in the encoder and the pooler is *gels*.

The most important parameters are the hidden size that we set to 768, a maximum sequence length of 512 (previously validated) so it is able to fit longer medical histories and then a batch size of 16 (adapted to the block size; the bigger the block size, the smaller the batch must be to fit the memory from the available resources). We train for a total number of 500 epochs and use only one of the text fields mentioned before, treatments or psychiatric history, but never both simultaneously. There is also a minimum number of visits allowed in the samples during training (removed in validation), which is set to 5, so the total quantity of patients in the training set decreases from $36,990$ to $16,972$ and from $9,248$ to $4,236$ in the validation set.

3.4. PsyBERT Imputing Missing Diagnoses

3.4.1. Context

The missing data scenario is an issue more and more frequent in data modeling nowadays. With so high quantities of data samples, it is very common losing information or, as in medical environments, directly not having some patient information due to lack of awareness. After analysing our dataset from psychiatry patients, we realized of this problematic with a missing rate of more than the half of the diagnoses in the EHR.

During Chapter 2 we also treated this task. More specifically, we tried to impute missing words or, more correctly, masked words with the help of a regularized Transformer, NoRBERT. Then, we proved that Transformers-based models are a good tool to explore in these scenarios. Now, in this chapter, we do also want to solve the missing values task. Nevertheless, instead of words, we impute diagnoses, following the same analogy we used to describe PyBERT with BERT.

3.4.2. Models

Training PsyBERT for imputing missing diagnoses means we stay in the first stage in the procedure (no finetuning is still applied), using a MLM policy and a mask rate of 0.15 as the original work of BERT with the main difference of masking diagnoses instead of words. Figure 3.4 shows a diagram of the *Sex-treatments PsyBERT* structure used to solve this task. Take into account that position and segment embeddings from Figure 3.3 are not included cause they are created by the model in the *embeddings* layer.

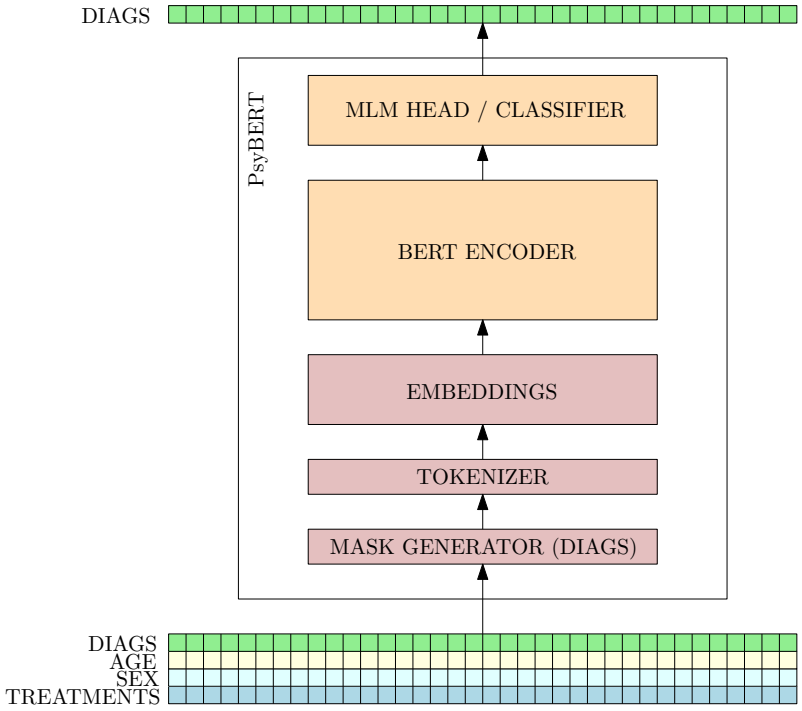


Figure 3.4: Diagram for the MLM pre-training in PsyBERT structure. Orange color means BERT blocks and brown color refers to modified and adaptations in the model. This is an example of the structure in *Sex-treatments PsyBERT*, so inputs in the model belong to diagnoses, age, sex and text about treatments from the patient visits' to the hospital. The output corresponds to the reconstruction of uniquely the diagnosis sequence after the masking process. Then, the mask generator is only applied to this sequence in the input.

During this section we perform firstly a training step trying different models and configurations (See Table 3.2) and measuring the APS in the validation set. Secondly, we apply a test step using the samples from the EHR which had missing values (that we removed in the dataset creation) to predict their diagnoses and next use the professional help of a psychiatrist to validate the model output.

Following the architecture and configuration described in the previous section, we train PsyBERT with the MLM policy and the complete samples (with no real missing data) from our database. We study 4 variations of the same model, regarding the embedding fields that we combine in the first layer of the architecture. See Table 3.2 for a comparative of all variations. The first model is the original vanilla BEHRT work, only with age and

diagnoses information. Then, in the second model we include the sex embedding from PsyBERT. Finally, in third and fourth models we integrate treatments and history fields respectively to the sex one.

As the baseline we wanted to compare our model with a well-known Transformer dealing only with text data and with no sequential information and no combination of simultaneous fields from the EHR. In other words, we use a dataset built with one of the two free text fields as input and the label with the diagnoses as the output as a traditional classification task in NLP. In table 3.2, we also include these baselines. For this, we also use a pre-trained model from the Transformers library, *roberta-base-biomedical-clinical-es*, that is the RoBERTa model pre-trained in a biomedical dataset in Spanish, so it can take advantage of additional information, making the baseline a higher spot to overcome. We do not use any pre-trained model in PsyBERT.

We train this baseline with a set of 252,486 and 63,122 samples for training and test respectively (note these samples refer to visits and not to patients). That is a quantity much bigger than the one in our sequential scenario since we have divided here each patient in several input samples, one per register. We train during 200 epochs and keep the best parameters regarding the APS. We use a maximum length in the text field of 200 tokens, a batch size of 32 and a learning rate of 10^{-5}.

	BEHRT base	**Sex**	**Treatments (text)**	**History (text)**	**Sequential data**
Baseline Treatments	No	No	Yes	No	No
Baseline History	No	No	No	Yes	No
Vanilla BEHRT	Yes	No	No	No	Yes
Sex PsyBERT	Yes	Yes	No	No	Yes
Sex-treatments PsyBERT	Yes	Yes	Yes	No	Yes
Sex-history PsyBERT	Yes	Yes	No	Yes	Yes

Table 3.2: Comparative of the baselines, BEHRT and PsyBERT variations with the embeddings included in the first layer. 'Sequential data' means we differentiate samples per patient (concatenating visits) or not, and we do it per register.

3.4.3. Results

We train the models for an enough quantity of epochs and then save the best one in a validation set on each scenario. We measure the performance with the value of the Average Precision Score (APS) in the validation set.

In Table 3.3 we present the results of the baseline and obtain the best model with a 0.8613 value in the APS. We need to claim that the best performance of the model is achieved using the history data instead of treatments, a difference regarding the results with PsyBERT. We think that this field has more useful general information when predicting only with one feature. Nevertheless, when using more features, the treatments field may give more extra information that serves as a complement as we will see later.

Model	Epochs	Accuracy	F_1 (micro)	F_1 (macro)	APS
Treatments	188	0.7205	0.8137	0.5341	**0.8248**
History	135	0.7526	0.8480	0.5935	**0.8613**

Table 3.3: Baseline results in a non-sequential set-up with a pretrained RoBERTa model in medical data.

Table 3.4 shows the training results for all model variations. From that table we can claim that including sex embedding does not improve the vanilla BEHRT on its own but if we also add the free text with the treatments, we are able to improve the APS over one point. These conclusions make sense since 'psychiatry history' field does contain general information about the patient situation but 'treatments' informs about the drugs and doses, what is directly related to the mental disorder in particular. In addition, the APS of 0.87483 is even higher than the baseline in the simpler (non-sequential) scenario from Table 3.3. With Table 3.4 information, we select the third model as the best one for the second stage in the validation.

Model	Epochs	Training loss	Test loss	Test APS
Vanilla BEHRT	320	0.01322	0.02709	0.86159
Sex PsyBERT	450	0.01252	0.02760	0.8613
Sex-treatments PsyBERT	157	0.009806	0.027328	**0.87483**
Sex-history PsyBERT	91	0.01543	0.03240	0.82764

Table 3.4: Train and test values for the best model on each scenario during the training of BEHRT and PsyBERT with a MLM policy.

Figure 3.5 indicates the validation APS for different masking rates on both models with a free text field (Sex-treatments and Sex-history PsyBERT) so we can intuit how it would behave in a real scenario regarding the complexity of missing data. The APS is lower than in the previous table because in validation we remove the filter of a minimum of 5 visits so the number of samples increase as well as the difficulty in their prediction since they can have shorter sequences.

Finally, we present the results in a real scenario with missing samples. This test consists on filling the original EHR database with missing diagnoses and validate the results with the help of an expert. For this validation we use data from 7,555 patients. Remember from Section 3.2 that it represents the 57.1% from the original database. Following Figure 3.5, a missing rate of 0.57 would give an APS of 0.65 approximately, so this is a reasonable value that we expect to obtain in a real scenario with such a high rate of missing values.

In Table 3.5 we present the validated results. After the model prediction, all outputs were analysed by a psychiatrist to label them in a scale with 5 possible values. 'Error' and 'No clinical information' correspond to cases without enough information to get into a conclusion. 'No clinical sense' are misdiagnosed patients according to expert judgement. Both 'With clinical sense' and 'complete agreement' are good results (encompasses as

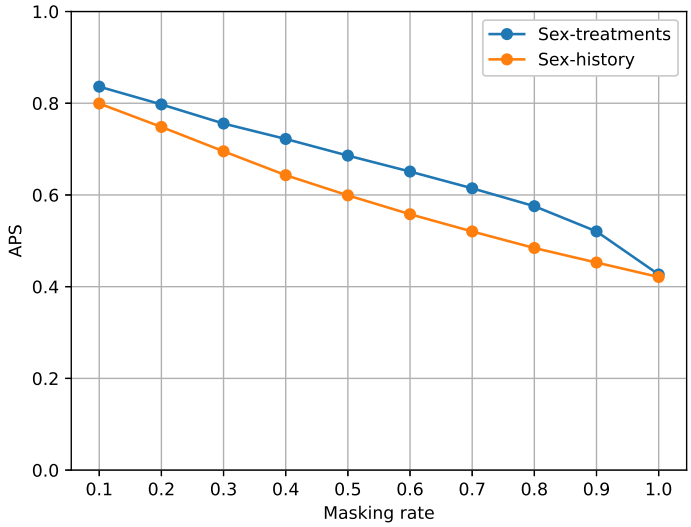


Figure 3.5: APS in the validation stage of the models with any text field and with different masking rates.

'Agreement' in Table 3.5). The former relates to cases where partial quantity of diagnoses were classified as positive, additional diagnoses from the expected ones were obtained or different choices from the ones the psychiatrist would have labeled are presented. However, all situations makes sense according to the expert. The latter refers to perfect diagnoses coincidence. Regarding this validation procedure, 67.42% of samples were correctly classified and 26.95% were wrong.

	Frequency	**Percent**
Error	20	0.26%
No clinical information	405	5.36%
No clinical sense	2036	26.95%
With clinical sense	1566	20.73%
Complete agreement	3528	46.7%
Agreement	5094	67.42%
No agreement	2036	26.95%

Table 3.5: Clinical validation in the prediction of missing diagnoses with a missing rate of 57.1% and a total of 768 different disorders.

3.4.4. Discussion

During this section we deal with a classification problem in a very complex scenario. The main difficulties we found are:

- The missing rate in the data is 57.1%.

- We try to classify 768 different disorders with only 16,972 patients in the training set.
- Comorbidity states are very often in psychiatry with up to seven disorders present simultaneously in a clinical episode.
- Misdiagnosed patients are also very frequent in this field of medicine.

Then, we have to point out that if we simplified the scenario, for example, by joining diagnoses from the same group and decreasing the number of classes, we could achieve much better results. However, we wanted to study the more complex option since the next section tries to specialize in a very concrete disorder among the whole set of classes, delusional disorder.

In addition, inside medicine, psychiatry is a branch very complex to be predicted in many cases. Its diagnosis sometimes requires long times of observation and the prevalence of a specific disorder may change very frequently, evolve to several diseases or turn out with another one constantly.

Taking into account all of these considerations and with the expertise opinion ahead, we can state that out results are very favorable. Therefore, we let a new field of study and open a research line to be improved by the inclusion of other fields in the EHR or through some architecture modifications. In any case, our results can help to clinicians as an external diagnosis advisor, or, as in the present work, as a tool to fill the missing data in the EHR.

3.5. PsyBERT detecting delusional patients

3.5.1. Context

Delusional disorder, also known as paranoia [126], constitutes one of the severe mental disorders. It is a psychiatric pathology whose main characteristic is the presence of well-systematized delusional ideas, without hallucinations or alterations of language or thought, which do not involve personality and cognitive impairment.

According to the DSM-V, delusional disorder has a lifetime prevalence of 0.02% [127]. However, prevalence for delusional disorder is much lower than other conditions like schizophrenia (1%) or mood disorders (5%); this may be in part due to underreporting of delusional disorder as those with delusional disorder may not seek mental health attention unless forced by family or friends.

Delusional disorder can have a major impact on an individual's ability to function, with difficulties in social and occupational functioning due to the well-systematized delusional ideas, and can even be difficult for family members to bear. Therefore, case identification

and early diagnosis is essential to establish an adequate treatment and try to avoid loss of patient functionality.

In other branches of medicine, diagnoses are usually based on the identification of underlying biological processes, whereas, in psychiatry, diagnoses are mainly based on the identification of symptoms throughout clinical evaluations in a cross-sectional and longitudinal manner. Delusional disorder is usually characterized by lack of illness awareness [128], which hinders outpatient follow-up and sometimes results in treatment drop-out. Thus, longitudinal assessment may be complex in patients suffering of delusional disorders which makes diagnosis even more complex. In addition to this, we must take into account that diagnostic stability in psychotic disorders, including delusional disorders, is generally low since diagnoses are difficult to establish and accordingly diagnostic error is high [129], [130].

Due to this problematic context, the task of predicting delusional patients get much more complex than any other disease. We must take into account that:

- The rate of positive labeled data is very low (around 2% as we present later).
- The are many misdiagnosed patients and missing information.
- The diagnosis usually lasts forever during patient's life.

With the model we describe in this section, we try to facilitate diagnosis in delusional disorders and promote early treatment, making the patient monitoring an easier process.

3.5.2. Models

To train our PsyBERT to detect delusional patients, we perform the same pre-training step with the MLM as in Section 3.4 with the model from Figure 3.4 and a posterior finetuning focused in our specific task (Figure 3.6). This task consists on a multi-label classification layer that learns to predict the diagnosis of the patient in the last visit to the clinician. Paranoid or delusional status is a disorder that once it is diagnosed, it stays forever in the patient's history. However, this is a behaviour that it is not well represented in the EHR data (its diagnosis is not maintained in the patient's registers) and therefore, it is difficult to be learnt by a machine learning model. Then, when the probability of developing this mental disorder is high enough, we consider the patient as a potential case that can be revised by a clinician.

With this scenario, we are interested in solving the problem of detecting false positive samples in the last visit. They are patients that are not diagnosed with any paranoid-related diagnosis, but might be suffering one of them. We finetune the model in a way that we obtain a probability output per diagnosis. Therefore, we need to set a threshold, ϵ, so we consider that a patient with a probability in paranoia higher than ϵ is not diagnosed. In resume, we train PsyBERT to predict the patient's diagnosis in the last visit and we only

keep the probability of the paranoid-related ones. Then we set a threshold and focus our results in the false positive samples, that will be the potential cases for delusional diagnosis that we have detected.

We keep the same train-validation split as before and find a rate of 2.29% and 2.37% of delusional patients in both sets respectively. It is a very low percentage of patients regarding what experts feel it should be. In terms of ICD-10 codification and with the help of a psychiatrist, we will consider during this work that diagnosis F60.0 and any F22 variant will be considered delusional. In our database, we find 5 different paranoid-related diagnosis, F60.0, F22, F22.0, F22.1 and F22.9, so we will treat the probability higher than the threshold in any of them as a potential delusional candidate.

3.5.3. Results

In the first place, we are describing a baseline we tried before studying more deeply the model described in Section 3.5.2. In this example, we use all diagnosis information from the EHR as input and train it in order to predict a binary classification label. This label is set to 1 in patients with any delusional diagnosis and 0 in the rest. After training this model we found what was expected in a so incorrectly labeled and very unbalanced dataset.

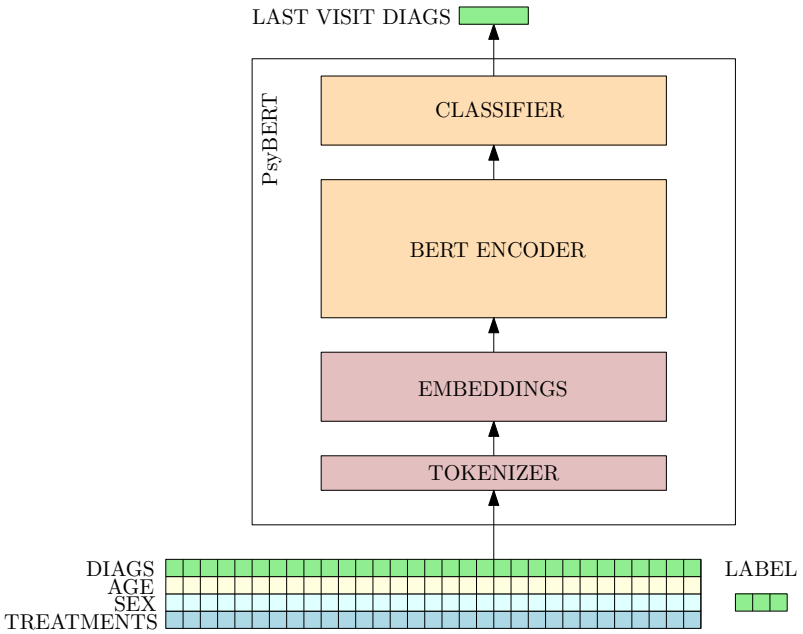
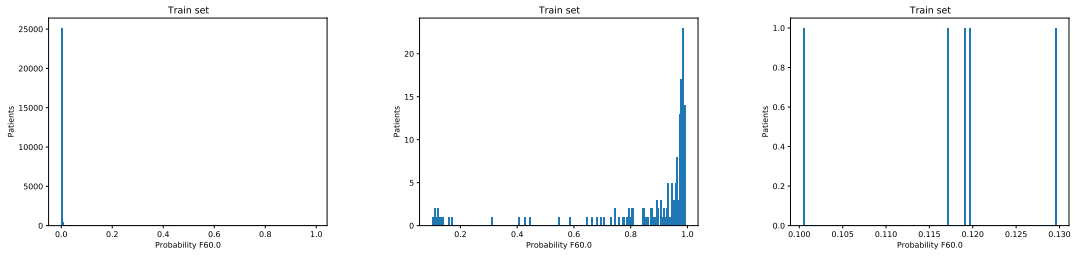


Figure 3.6: Diagram for the last visit classification fine-tuning in PsyBERT structure. Orange color means BERT blocks and brown color refers to modified and adaptations in the model. This is an example of the structure in *Sex-treatments PsyBERT*, so inputs in the model belong to diagnoses, age, sex and text about treatments from the patient visits' to the hospital. Label refers to the classification task, that is the diagnoses in the last visit, not shown in the other input sequences. The output corresponds to the probabilities for each diagnosis in last visit.

We obtain an APS of 0.99976 and a ROC of 0.99972, what means that this model works too well fitting the labels and it is no able to search non diagnosed delusional patients. One might think that this ROC reflects the perfect result. However, we need to remember that we are facing a problem with wrong or incomplete labels. Hence we do not find the highest precision and recall values. In this baseline, the probability histograms in delusional diagnoses show that the output result is very sure positive or very sure negative, but we do not obtain any patient with a doubtful probability where we could apply any threshold.

Due to the inability of the traditional approaches to deal with our classification problem as shown in the previous baseline, we present now the definite architecture results we obtain in solving this problem. The output obtained from the model after the finetuning of the classification layer consist on a probability value for each of the diagnosis in the last visit of the patient. As we are interested in delusional disorder, we only take this value and apply a threshold ϵ of 0.1 or 0.01 as the probability to consider the candidate as potential delusional.

Before setting this hyperparameter to any value, we printed and studied the histograms of the probability values in these diagnosis. By this procedure, we found out that we needed a very low threshold if we wanted to detect a significant quantity of delusional patients. We study these conclusions in Figure 3.7 for an example with the training samples and the F60 diagnosis. Firstly, Figure 3.7a shows the whole histogram where we can confirm that the vast majority of patients are given a probability of 0 in F60. If we zoom this histogram by cutting probabilities below 0.1 (Figure 3.7b), we see a peak in probability 1 and close to it, but we also appreciate some samples spread out all the axe. Finally, in Figure 3.7c we repeat the same graph but showing only non diagnosed samples. In that scenario is where we appreciate some possible candidates to be delusional, all of the with low probability. These figures reflect the results of the first case, using vanilla BEHRT but give us an idea of what can be obtained.



(a) (b) (c)

Figure 3.7: Histogram of probabilities in F60 diagnosis from training samples (a) in general, (b) with a zoom in values higher than 0.1 and (c) with the same zoom but only non diagnosed patients.

As in Section 3.4, we try four different models, BEHRT as the vanilla original work, and three more adding the embbedings mentioned before: vanilla, PsyBERT with sex

embedding, PsyBERT with sex embedding and treatments embedding and, PsyBERT with sex embedding and psychiatric history embedding. We pre-train the models during 300 epochs and finetune in the classification task for 400 epochs. The scores obtained in all cases are presented in Table 3.6. We can only compare them in pairs since the two first ones have a block length of 64 and the others 256.

Model	APS	ROC	Loss
Vanilla BEHRT	0.9365	0.9918	0.0405
Sex PsyBERT	0.9325	0.9905	0.0433
Sex-treatments PsyBERT*	0.9484	0.9895	0.2193
Sex-history PsyBERT*	0.9479	0.9903	0.1999

Table 3.6: APS, ROC and loss after finetuning the 4 variants of the model. *256 block length versus 64 the other 2 variants.

Following the procedure described before, the number of false positive samples obtained by the models is presented in Table 3.7. While in vanilla and sex-case it is fine to use $\epsilon = 0.1$, it is not suitable in the variants with free text since it is not able to detect any possible delusional. For those variants we apply $\epsilon = 0.01$ and obtain a understandable quantity of false positives. However, that threshold in the other models shoot up the candidates to around 500, what is too high. The bold numbers in the table correspond with the potential cases of delusional patients that we select. Due to the low prevalence of this disorders, they seem reasonable numbers before doing any validation.

Model	$\epsilon = 0.1$	$\epsilon = 0.01$
Vanilla BEHRT	**13**	565
Sex PsyBERT	**26**	498
Sex-treatments PsyBERT	0	**2**
Sex-history PsyBERT	0	**22**

Table 3.7: False positives in all models. Bold values correspond to the potential patients with delusional disorder or paranoia.

Next, we include in Table 3.8 some measures of diagnostic accuracy regarding sensitivity and specificity values[10]. In next table and equations, FP refers to false positive, TP true positive, FN false negative and TN true negative samples. The **sensitivity** is defined as the probability of a positive test result given disease (Equation 3.1a) and the **specificity** is the probability of a negative test result given non-disease (Equation 3.1b). We also include the values for the probability of disease given positive test result (PPV) detailed in Equation 3.1c and the probability of non-disease given negative test result (NPV) in Equation 3.1d.

[10]https://www.acomed-statistik.de/

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.1a}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3.1b}$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.1c}$$

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{3.1d}$$

Results in Table 3.8 complete the study about the accuracy of our model and its variations so interpretation is remained. All percentages in the measures are above 85%, what indicates a good performance in the model.

	Vanilla BEHRT	**Sex PsyBERT**	**Sex-treatments PsyBERT**	**Sex-history PsyBERT**
FP	13	26	2	22
TP	782	788	777	773
FN	114	108	119	123
TN	31663	31650	31674	31654
Sensitivity	87.28% (84.92-89.39%)	87.95% (85.63-90.01%)	86.72% (84.32-88.87%)	86.27% (83.84-88.46%)
Specificity	99.96% (99.93-99.98%)	99.92% (99.88-99.95%)	99.99% (99.98-100%))	99.93% (99.89-99.96%)
PPV	98.36% (97.22-99.13%)	96.81% (95.35-97.9%)	99.74% (99.08-99.97%)	97.23% (95.84-98.26%)
NPV	99.64% (99.57-99.7%)	99.66% (99.59-99.72%)	99.63% (99.55-99.69%)	99.61% (99.54-99.68%)

Table 3.8: Sensitivity an specificity table. Parenthesis values correspond to confidence intervals regarding a width of 95%. The disease prevalence is 2.8%, correctly reflected by the study.

In order to check the reliability of our model, we pursue a final step of interpretation with the validation of some clinicians. According to the potential cases in Table 3.7, we perform the following analysis: we take the clinic history of these patients and analyse the usual disorders that these individuals are diagnosed. These results are presented in Table 3.9. After this detailed study, we concluded that the potential positive samples belong to mainly patients with disorders due to use of alcohol, tobacco or cannabinoids, unstable or unspecified personality disorders, paranoid schizophrenia, psychosis and adjustment disorders. Regarding ICD-10 specifications[11], all these disorders are related to emotional or perceptual disturbances that interferes social functioning, delusions, hallucinations, paranoia, delirium, manic episodes or impulsive acts among other symptoms. All of them are directly related to delusional disorder, so we can claim the effective results obtained by our model.

3.5.4. Discussion

During this section we present a diagnostic tool with many advantages in today's treatment of mental health. We use a Transformer-based architecture to predict the probability

[11] https://icd.who.int/browse10/2019/en

Model	Diagnoses
Vanilla BEHRT	Disorders due to use of **alcohol**, disorders due to use of **tobacco**, emotionally **unstable** personality disorder, mental disorders due to **brain damage** and dysfunction, emotionally **unstable** personality disorder, **adjustment disorders**, unspecified personality disorder, schizophrenia, anxiety, **hyperkinetic** disorders, **psychosis**
Sex PsyBERT	Disorders due to use of **alcohol**, disorders due to the use of **cannabinoids**, unspecified personality disorder, **psychosis**, anxiety, **paranoid schizophrenia**, dysthymia, **schizotypal** disorder, **bipolar** affective disorders, agoraphobia, mental disorders due to **brain damage** and dysfunction
Sex-treatments PsyBERT	**Bipolar** affective disorder, mania, **psychosis**
Sex-history PsyBERT	Disorders due to use of **alcohol**, disorders due to the use of **cannabinoids**, **paranoid schizophrenia**, **adjustment disorders**, unspecified personality disorder, delirium, dysthymia, emotionally **unstable** personality disorder, **hyperkinetic** disorders, **bipolar** affective disorder, anankastic personality disorder, acute and transient **psychotic** disorders, **schizotypal** disorder

Table 3.9: Diagnoses from potential cases of delusional disorder. Common states are highlighted in bold.

of delusional disorder in patients from mental disorders and we take this output to find potential candidates of suffering the pathology. We analyse the results and agree in other related disorders that prevalence in these potential cases such as paranoid schizophrenia, psychosis or the consume of alcohol.

As a conclusion, we enumerate some of the advantages that these model has:

1. It allows the evidence of a diagnosis or the detection of an error in the diagnoses established by the clinician.
2. It imputes the codification of diagnoses in data bases without this information in order to unify the nomenclature.
3. It helps to set a diagnosis in a more reliable way when the patient has provisional diagnoses, especially in first episodes.
4. It could shorten or eliminate completely the timing to decide a diagnosis since many diagnoses need a temporal criterion to express the symptoms (e.g. at least six months in schizophrenia patients).

4. PROBABILISTIC METHODS IN MENTAL E-HEALTH QUESTIONNAIRES

Introduction

The development of new technologies shows promise for causing a revolution in the way chronic diseases are followed and treated [131]. In the past few decades, we have seen the following two major technological changes directly connected to the availability of medical information: (1) introduction of electronic health records in most health care facilities, and (2) accessibility to portable devices capable of acquiring information about their users. Both systems are already being used to enhance communication between health providers and final users and to improve the overall performance of health care. Indeed, public and private entities are massively investing in the development of web-based platforms or smartphone apps through which patients can organize their medical agenda, have access to all or part of their medical records, provide their input, and join their medical referents [132].

It seems reasonable to believe that the follow-up of persons with mental illness will be improved if eHealth systems lead to an increased interaction with health care providers. Our group and others have shown that electronic assessment is feasible with proper adaptation [133]. Efficient monitoring may prompt health responses in cases of emergency [134], inform accurately about real-life behaviors between medical appointments, reduce unnecessary visits, and sustain therapeutic decisions [135]. Other parameters, such as biomarkers and input from close relatives, can be added to the monitoring system. This kind of ecosystem already exists [136], and a growing body of evidence has shown that eHealth tools improve treatment outcomes in terms of engagement, symptom improvement, well-being, and self-care [137]–[143]. Their combination with machine learning techniques has also shown positive impacts on the diagnosis, prediction, and prevention of several diseases, such as cancer [144]–[146].

Despite these advances, the majority of potential users, such as elderly persons with low educational levels [147], seem to be unenthusiastic about e-mental health tools [148]. Utilization rates have been associated with the characteristics of health professionals [149], but there is little knowledge about the kinds of patients who use e-mental health, how they become users, and what are their patterns of use. Young age, high education, and dissatisfaction with the health care system might be common features among eHealth users [150]. We do not know if this profile also applies to patients with mental disorders, but their digital phenotype [151] is likely to contain valuable information for clinicians and providers alike [152], [153]. Alterations in the patterns of use could help clinicians to detect pathological or risky behaviors and individual needs, and increase treatment efficiency.

Smartphones are especially apt for monitoring symptoms, given their increasing ubiquity, their versatility, and the users' widespread habit of carrying them at all times [154], [155]. Smartphone monitoring is being increasingly used in biomedical research. There are two modalities of smartphone monitoring: passive —by using the smartphone's native sensors—, and active —by asking questions to participants via the device, resulting in a real-time and takes place in the participant's usual environment. Smartphone monitoring reduces recall bias, respect ecological validity, and facilitate the simultaneous collection of data [156], [157].

In these works, a nonparametric latent feature model based on the Indian Buffet Process (IBP) explores the response patterns of psychiatric outpatients to different web-based questionnaires. These questionnaires are asked via the smartphone in different ways. In the first model (Section 4.1), we take one sample per patient, which belongs to the first time the patient or user answered the questionnaire. So, all questions were asked at the same instant. For the other two works (Sections 4.2 and 4.3), this mechanism was improved and only sets of few questions are asked in different instants during the day. So, the frequency of the questions was particular regarding the importance of each item. Furthermore, this algorithm deals the behaviour that the user may get bored of answering every day the same and all of the questions. Finally, another difference in the second and third works is the inclusion of the time evolution, that will help us apply the model for different application as the suicide study before and during COVID quarantine in Section 4.3. Contrary to the first work where each patient correspond to a determined profile, in these cases, each patient may belong to different profiles along the time evolution.

Following, in this chapter we develop the three works in different sections and we going towards less detail in the model description but more complex and precise application.

4.1. Psychiatric Profiles of eHealth Users

4.1.1. Objectives

In this first work, I look for a general study without focusing in patients with specific features but defining the different profiles according to response patterns in a health questionnaire. Then, I link the profiles with psychiatric disorders regarding the EHR of each individual. The study thus specifically describes patients with psychiatric diagnoses who have used an eHealth application at least one, what ends with a total of 2254 patients. The goal of using a probabilistic technique such as IBP is to associate information from different questionnaires and assessments in a plausible model that could serve ultimately to plan health care delivery.

4.1.2. Data

Participants were recruited from psychiatric outpatient facilities in the catchment area of *Fundación Jiménez Díaz*, a University Hospital in Madrid, Spain. This hospital is part of the National Health Service and provides medical coverage to about 850,000 people. From May 2014 onwards, all clinicians working at the six mental health centers of the catchment area received specific training and were encouraged to use the **MEmind Wellness Tracker** systematically in their clinical activity. The MEmind application is apt for both Android and iOS operating systems and is freely downloadable and available in App Store and Google Play. However, for its activation it is necessary to have a username and password, which are provided by the researcher at the baseline visit. A detailed description of the MEmind app has been published elsewhere [136], [158]. A total of 2254 patients signed up on the MEmind platform and completed the assessment, and they were subsequently included in the study. The assessment comprised the collection of information about sociodemographic features and diagnoses. Participants also filled up a short questionnaire. For this study, we used broad inclusion criteria. Every patient attending psychiatric consultations independent of diagnosis was considered. Thus, all clinicians in the catchment area were instructed to propose the use of the web application to every outpatient they saw with no restriction whatsoever regarding their diagnoses or their clinical statuses. The total number of outpatients who consulted during the study period was 30808.

The application also has a free text field in which the patient can write comments about their state. All answers were stored in the device and when the mobile phone connected to a WIFI network, they were uploaded to a secure web server accessible by clinicians and researchers (see data protection section below).

For the purpose of this work, we included only participants who voluntarily accessed the application and responded to the open-text field. We made this choice to select proactive participants who completed most of the questions at the user end. We noted a missing data rate of 12%, which resulted from the sum of clinical missing data and a lack of completeness of the questionnaires at the user end of the application.

This study was performed in agreement with the ethic requirements of the Declaration of Helsinki (World Medical Association, 2013) and was approved by the Institutional Review Board of the University Hospital *Fundación Jiménez Díaz* (Madrid, Spain). All participants provided written informed consent to participate in the study.

Questionnaires

The data set consists of 23 questions from the following three different questionnaires: (1) a brief day assessment related to sleep quality, appetite, medication intake, aggressiveness, and suicidal behavior (six items); (2) the Who-5 Well-Being Index [159] (five items); and (3) the ninth version of the General Health Questionnaire [160] (12 items). All these

questionnaires are short self-reported measures of current mental well-being. All items are yes-or-no questions, followed by the degree of agreement reported on a Likert scale (0 to 100 points). Although participants could repeat the assessment, only data from the first report was included in the model.

Clinical Diagnoses

Diagnostic coding was based on the ICD. Thus, diagnoses of mental disorders were classified into 10 groups (F0 to F9) according to ICD-10 (See Appendix A for codification description of the diagnoses). The corresponding physician coded the diagnosis for each patient and completed the Clinical Global Impression (**CGI**) scale [161], which reflects the global functioning of a patient according to the view of the clinician on a scale (0 to 7 points). The CGI scale provides a summary measure accounting for patient history, psychosocial factors, behavior, and the impact of symptoms on the patient's ability to function (See Appendix B).

Data protection

Data were stored in a secure external server created for research purposes. Only the principal investigator in the psychiatric side (Enrique Baca García) had an access code to the server. MEmind used AES-256 algorithms and 256-bit keys to encrypt data. Keys were protected by a professional key management infrastructure, which implemented strong logical and physical security controls to prevent unauthorized access. An external auditor guaranteed that security measures met the Organic Law for Data Protection standards at a high protection level.

4.1.3. Methods

Data processing

First, the scores for the items with a positive valence in the questionnaire data set (items 1 to 15) were inversed. In this way, a higher score for any item of the questionnaire indicated poorer mental health. Second, we dichotomized every item score using a specific threshold in order to code the top 10% scores with the value "1" and the remaining 90% with the value "0" as a model criterion justified in the model section below (See Table 4.1 for threshold details).

The use of a centesimal scale increases the sensibility of the questionnaire. Responders tend to avoid extreme values unless they identify completely with them [162], but extreme responders do not seem to be affected by the length of the response scale [163]. By using the highest scores, we made sure that only the extreme responders were separated. The histograms of scores before being dichotomized are shown in Figure 4.1.

Item	Percentage of '1' values	Threshold
1	9.94	0.62
2	9.75	0.74
3	9.67	0.78
4	11.22	0.9
5	23.51	0.9
6	46.49	0.9
7	9.93	0.83
8	9.76	0.87
9	10.16	0.9
10	15.31	0.9
11	9.71	0.88
12	9.89	0.72
13	13.43	0.9
14	9.85	0.9
15	9.63	0.84
16	9.98	0.85
17	9.94	0.9
18	9.72	0.89
19	9.4	0.87
20	11.49	0.9
21	15.7	0.9
22	29.06	0.9
23	9.58	0.9

Table 4.1: Dichotomy process of data from the questionnaire scores

The clinical records of the participants provided a second source of data. These records included sex, age, clinical diagnoses, and CGI values. CGI values presented missing data, so subsequent analyses including this variable were carried out with a total sample of 2000 participants. For analyses involving clinical diagnoses, the total sample was 1787 participants. All patients with missing data were excluded from this part of the data modeling. Comorbid diagnoses were also examined when present.

Model

We applied the **Sparse Poisson Factorization Model** (SPFM) to fit the data. The SPFM [22] is based on the IBP [20], [21], a non-parametric probabilistic method that proposes a sparse decomposition of the variables. Non-parametric Bayesian techniques are frequently employed in machine learning in order to discover the internal structure of a dataset by modelling the underlying correlations among the given variables. In the SPFM, the

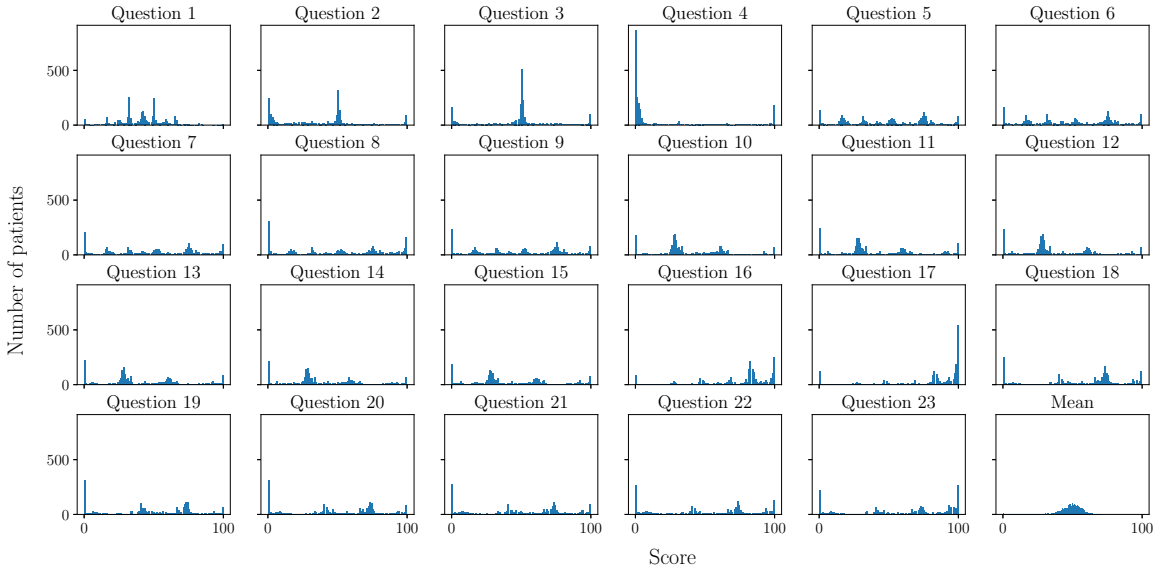


Figure 4.1: Histograms of scores provided by eHealth users to each of the 23 questions of the mental well-being questionnaire. Scores range from 0 to 100. The last histogram presents the average score for all questions.

dimension in the latent structure of the data, K, is unknown and learnt during the training. The word 'sparsity' describes a way of considering only a little subset of the model coefficients different to zero. As a condition, the input data must be binary or categorical and for a better convergence the number of non-zero elements must be minimum.

The SPFM decomposes the input matrix into the following two non-negative and disperse matrixes, Z and B, what makes the latent features easy to calculate and interpret. The factorization is detailed in the set of equations 4.1, with the definition of the priors and data distributions.

$$\begin{aligned} X_{nd} &\sim \text{Poisson}\left(Z_n \cdot B_d\right) \\ Z &\sim \text{IBP}\left(\alpha\right) \\ B_{kd} &\sim \text{Gamma}\left(\alpha_B, \frac{\mu_B}{\alpha_B}\right) \end{aligned} \tag{4.1}$$

In the previous equations, n index indicates the sample and d the feature of the data, that is, each of the questions in our scenario. The binary Z matrix represents the number of active features for each patient as shown in Figure 4.2. The B prior follows a Gamma distribution with α_B controlling the scale and μ_B the mean. As a result, it weights the contribution of each feature to each item of the questionnaire. Each feature is characterized by precise values on the 23 questions. A higher weight (B) of a feature for an item is associated with a greater probability to find a high score in that item when that particular feature is active. The SPFM also estimates a bias term, a feature that is presented in all the patients of the sample. The bias term is the "default" situation of an eHealth user and

represents a profile shared by all patients that is independent of any additional feature. In that sense, the bias term allows the algorithm to be shifted to better fit the data.

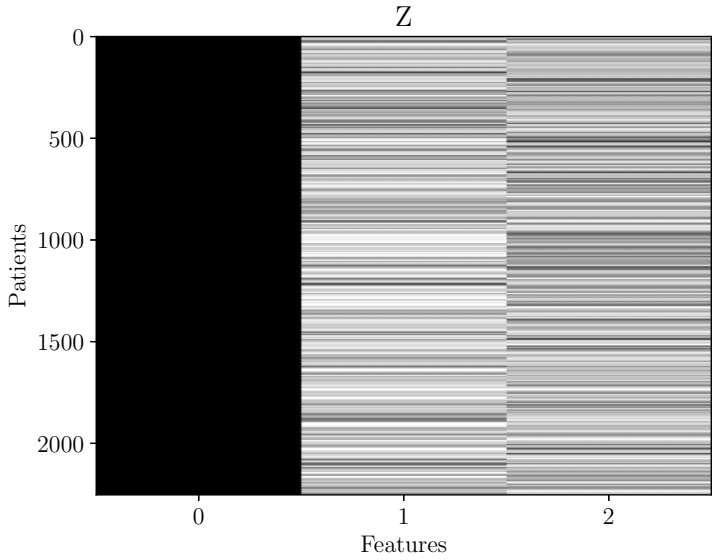
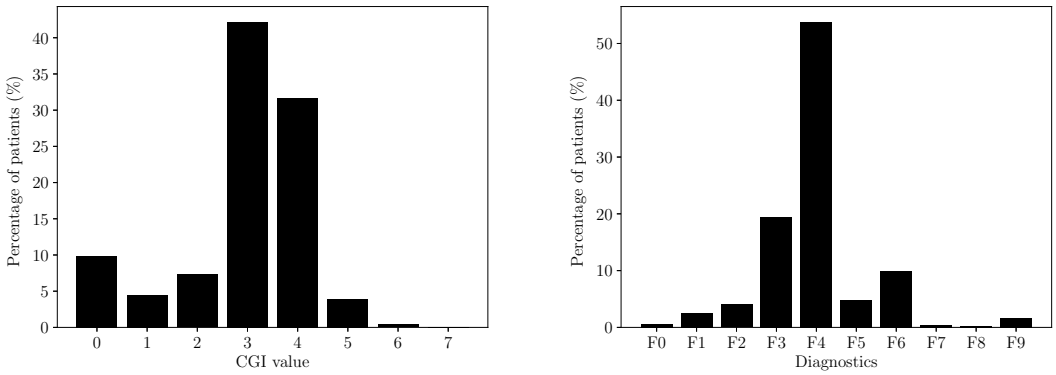


Figure 4.2: The binary Z matrix, presenting the number of active features for each patient. Each line corresponds to a single patient. All patients present the bias term or feature 0.

After applying the SPFM, we use a basic clustering method, K-means [164], to obtain different profiles. The procedure classifies a given data set through a certain number of clusters fixed in advance. This method, applied on the Z matrix, associates data with similar characteristics into different clusters by using centroids. Thus, it allows clustering patients who show similar activation of their features. For example, in our study we obtain 3 features as shown in figure 4.2, omitting the bias always active, a patient could have only feature 1, only feature 2, both features activated or none of them in his or her answers to the questionnaire. This turns out in 4 different profiles as we will discuss in Section 4.1.4.

4.1.4. Results

The sample involved 2254 patients, including 1184 (52.53%) women, 795 (35.27%) men, and 275 (12.20%) patients with missing data on sex. The mean age was 52.0 years (SD 15.1). Medical reports about the patients showed a CGI mean score of 2.95 (SD 1.95) with a high percentage of participants scoring 3 (mildly ill; 844/2000, 42.20%) or 4 (moderately ill; 632/2000, 31.60%) and only a few scoring 6 or 7 (severely ill or extremely ill; 10/2000, 0.50%). The CGI distribution in the sample is presented in Figure 4.3a. According to the ICD-10 criteria and Figure 4.3b, participants with mood disorders (F3; 347/1787, 19.43%), stress-related, neurotic and somatoform disorders (F4; 962/1787, 53.82%), and adult personality disorders (F6; 178/1787, 9.96%) represented most of the sample. F0-F9 codes represent main ICD-10 diagnostic categories for psychiatric disorders.



(a) Percentage of patients regarding CGI scores. (b) Percentage of patients according to ICD-10.

Figure 4.3

Model output

The SPFM latent model analysis found the following three components in the assessment: one bias term and two features, as exposed in Figure 4.2. Both the bias term and features involved groups of items of the questionnaire that are particularly informative. The bias term is present for all patients and reflects a common behavioral pattern. On the other hand, features 1 and 2 are based on subsets of answers with high informational value to discriminate patients. Features 1 and 2 can be present or absent for a particular patient.

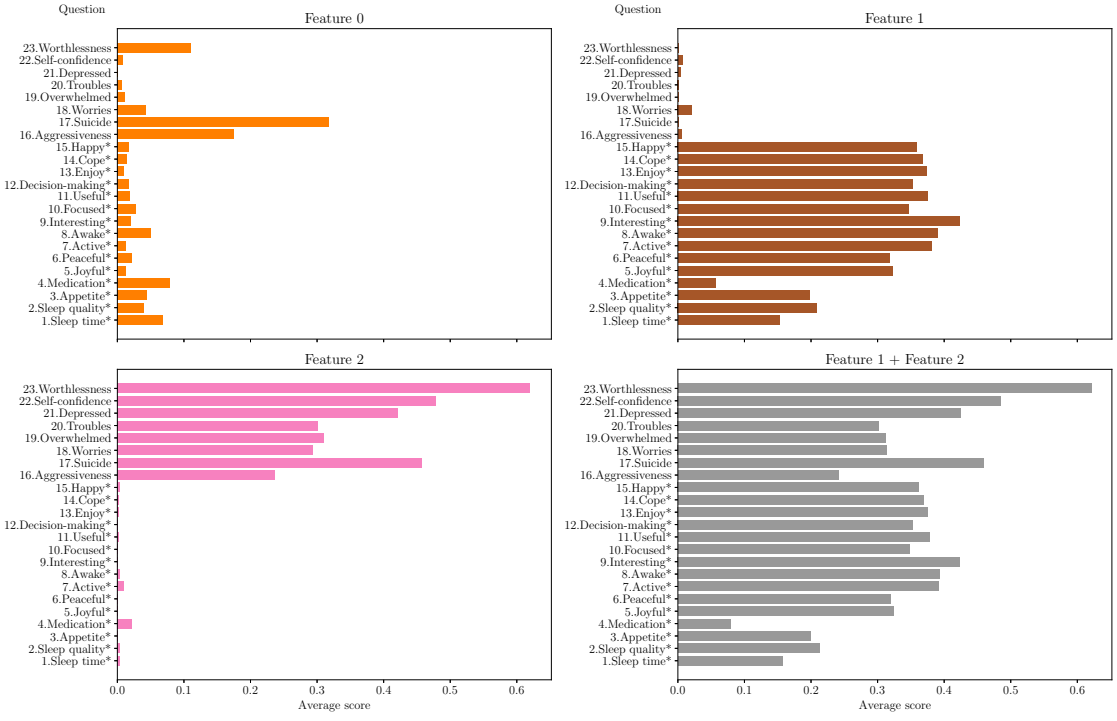


Figure 4.4: Average scores for each item of the self-reported questionnaire of current mental well-being according to the B matrix.

They are present if the corresponding subset of responses has a high score and are absent if the corresponding subset of responses has a low score. In other words, a higher weight in B of a feature for an item is associated with a greater probability to find a high score in that item when that particular feature is active. This correspondence between features and questions is exposed in Figure 4.4.

Questions	Profile 0 Bias term (0)	Profile 1 Feature 0+1	Profile 2 Feature 0+2	Profile 3 Feature 0+1+2
1. How many hours did you sleep today? (from 0 to 12)[12]	0.0689	0.2215	0.0731	0.2256
2. Quality of sleep[12]	0.0390	0.2480	0.0430	0.2520
3. Do you have appetite?[12]	0.0446	0.2426	0.0455	0.2436
4. Do you take your medication?[12]	0.0795	0.1361	0.1020	0.1587
5. I felt joyful and with good mood[12]	0.0133	0.3362	0.0140	0.3369
6. I felt peaceful and relaxed[12]	0.0223	0.3404	0.0235	0.3416
7. I felt active and robust[12]	0.0130	0.3941	0.0234	0.4045
8. I felt awake, fresh, and rested[12]	0.0496	0.4391	0.0538	0.4433
9. My daily life has many interesting things[12]	0.0199	0.4435	0.0199	0.4435
10. Have you been able to keep focus on the tasks you did?[12]	0.0276	0.3743	0.0289	0.3757
11. Have you felt that you have a useful role in life?[12]	0.0187	0.3944	0.0206	0.3962
12. Have you felt able to make decisions?[12]	0.0173	0.3702	0.0175	0.3705
13. Have you enjoyed regular activities from daily life?[12]	0.0100	0.3839	0.0116	0.3856
14. Have you felt able to cope with your issues?[12]	0.0149	0.3824	0.0166	0.3840
15. Do you feel reasonably happy taking into account the circumstances?[12]	0.0174	0.3765	0.0205	0.3796
16. Do you feel aggressiveness?	0.1745	0.1798	0.4107	0.4160
17. Do you have suicidal thoughts?	0.3177	0.3189	0.7750	0.7762
18. Have you had worries interfering with your sleep?	0.0422	0.0633	0.3354	0.3565
19. Have you felt constantly overwhelmed or tense?	0.0109	0.0116	0.3217	0.3224
20. Have you felt unable to overcome your troubles?	0.0060	0.0076	0.3066	0.3081
21. Have you felt unhappy or depressed?	0.0001	0.0037	0.4215	0.4253
22. Have you lost self-confidence?	0.0083	0.0154	0.4867	0.4939
23. Have you felt worthlessness?	0.1097	0.1107	0.7296	0.7306

Table 4.2: This table shows the average score for each item in the self-reported questionnaire of current mental well-being according to B. A question score on each profile is the sum of its B values from the corresponding features (suitable columns in the matrix).

The K-means algorithm applied on the Z matrix established four different patient profiles according to the presence of none, one, or both features. Profile 0 presents only the bias term, profile 1 presents the bias term plus feature 1, profile 2 presents the bias term plus feature 2, and profile 3 presents the bias term and each feature. The number of patients in each profile is shown in Table 4.3. In Figure 4.4 as well as in Table 4.2, we can appreciate how each profile responds differently to the questionnaire. Due to the negative connotation in the items from the questionnaire, the highest the values in that figure means the worse is the mental state on the patient, being the profile 3 with all features active, the most critical case in the scenario. Moreover, in Table 4.2, it is detailed the numbers obtained by the model and the literal description of each question.

[12]The scores from these items were reversed during data processing.

Profile	Number of patients	Feature
0	1113	Bias term (0)
1	480	0+1
2	616	0+2
3	45	0+1+2

Table 4.3: Number of patients and feature distribution for each profile.

Analysis

In this section, we are interpreting the model output with expertise help from the medical sector. The bias term was associated with high scores in suicide thoughts and aggressiveness (items 16 and 17), as well as feelings of worthlessness (item 23). All patients in our sample shared the bias term, but about half of them (n=1141) also presented one or two different features. Those presenting feature 1 were included in profile 1, which was characterized by the absence of positive mood, low sleep quality, low energy, and feelings of loss of control (items 1-3 and 5-15). Patients presenting feature 2 were included in profile 2, which was characterized by intense suicidal thoughts, aggressiveness, intense feelings of depression and worthlessness, low self-confidence, and worries interfering with sleep (items 16-23). All these characteristics were simultaneously active in patients from profile 3, who presented simultaneously both features. No statistical differences were found between the profiles regarding the distribution of age or sex ($F_3 = 1.391, P = 0.24$ and $\chi^2_3 = 0.56, P = 0.90$).

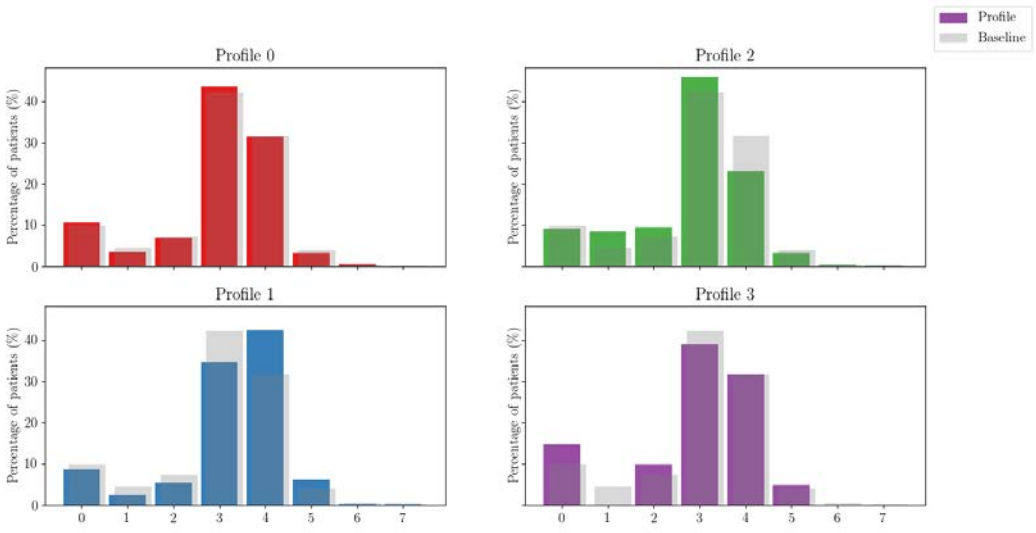


Figure 4.5: CGI for patients on each profile

After modeling the data, we compared CGI scores and clinical diagnoses between profiles. The results showed that the CGI scores were higher than the mean in profile 1 (3.2, SD 1.27), with the largest percentage of participants evaluated with a score of 4 (192/453,

42.4%). For profile 2, the CGI scores were lower than the mean (2.78, SD 1.28), with a high percentage of participants evaluated with a score of 3 (256/557, 45.9%). Results in profile 3 were not compared given the low number of patients. Figure 4.5 collects this information.

Most diagnoses fell within the F4, F3, and F6 categories in each profile and in the total sample, corresponding with affective disorders, neurotic and stress-related disorders, and disorders of adult personality and behavior. The distribution of participants with profile 1 was similar for all the types of diagnoses. However, profile 2 seemed to be more frequent among patients with diagnoses of schizophrenia and psychological, behavioral, and emotional disorders with onset in childhood/adolescence (F2: 43/90, 48%; F8: 3/5, 60%; and F9: 14/38, 37%; see Figure 4.6 and Table 4.4 for more details).

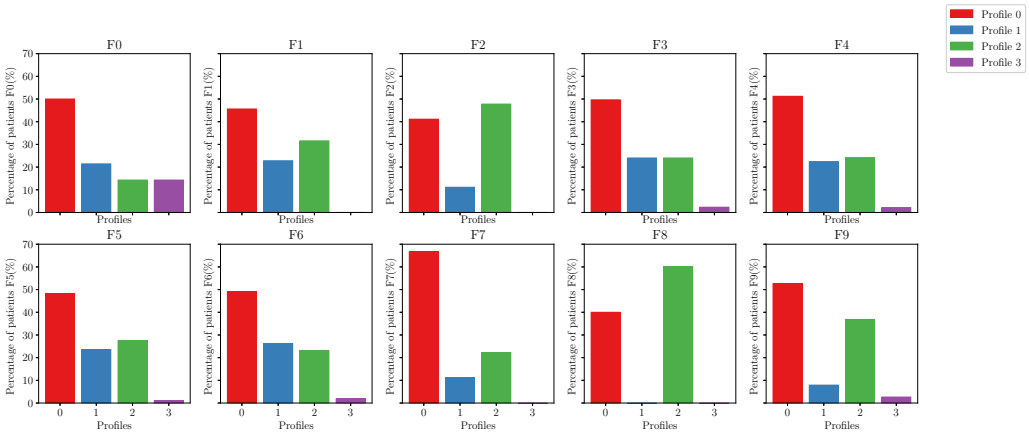


Figure 4.6: Distribution of profiles for patients from each disorder regarding the ICD-10 categories.

Category	Profile, n(%)			
	0	1	2	3
F0	7 (50.0)	3 (21.4)	2 (14.3)	2 (14.3)
F1	26 (45.6)	13 (22.8)	18 (31.6)	0 (0.0)
F2	37 (41.1)	10 (11.1)	43 (47.8)	0 (0.0)
F3	215 (49.6)	104 (24.0)	104 (24.0)	10 (2.3)
F4	614 (51.2)	269 (22.4)	290 (24.2)	26 (2.2)
F5	51 (48.1)	25 (23.6)	29 (27.4)	1 (0.9)
F6	109 (49.1)	58 (26.1)	51 (23.0)	4 (1.8)
F7	6 (66.6)	1 (11.1)	2 (22.2)	0 (0.0)
F8	2 (40.0)	0 (0.0)	3 (60.0)	0 (0.0)
F9	20 (52.6)	3 (7.9)	14 (36.8)	1 (2.6)

Table 4.4: This table shows the distribution of patient profiles according to the main ICD-10 diagnostic categories for psychiatric disorders (F0-F9).

4.1.5. Discussion

The data modeling approach we applied was able to discriminate four different profiles of patients based on the answers to a brief electronic questionnaire. All profiles shared a component associated with feelings of aggressiveness, worthlessness, and suicidal thoughts (bias term or profile 0), which seemed to be common among patients who used e-mental health tools [158], such as the **MEmind Wellness Tracker**.

Sex and age distributions showed very little variability across the profiles, facilitating comparisons between them. In addition to profile 0 (bias term, default pattern), three profiles were found based on the scores of different sets of questions. It is important to bear in mind that a feature was classified as active only when the scores were in the top 10% of the corresponding items. For example, even if the item of low sleep quality is absent from profile 2, a patient with that profile could still have high scores in that item compared with the general population and thus have relatively low sleep quality.

Patients in profile 1 reported a lack of positive mood, low quality of sleep, low energy, feelings of loss of control, and difficulties to face problems. These symptoms could be reactive to life difficulties and partly due to a lack of coping skills. Patients in profile 2 presented high scores in depressive feelings, worries interfering with their sleep, feelings of being overwhelmed and unable to overcome troubles, low self-confidence, and feelings of worthlessness. This pattern seems to be related with a greater inward focus and depressive-like symptomatology. Interestingly, patients in profile 2 also reported the highest scores for suicidal thoughts and feelings of aggressiveness. Indeed, patients in profile 2 reported five of the 10 ICD-10 diagnostic criteria for a depressive episode, including disturbed sleep, depressive feelings, reduced self-confidence, ideas of worthlessness, and ideas of suicide [125]. Surprisingly, those in profile 2 were evaluated by their physicians as having a higher level of functionality (CGI) than those in profile 1, despite higher levels of suicidal thoughts, aggressiveness, and depression in profile 2. This points to discordances between the medical assessment and the self-reported momentary assessment. Finally, profile 3 involved a small group of patients with high scores in all the items of the questionnaire. They shared the features of profile 1 and profile 2, and reported the most severely affected psychological state in our sample (the highest levels of distress).

Our study suggests that the analysis of data from electronic self-assessments can discriminate profiles or clusters of patients sharing similar clinical characteristics. These features do not seem to overlap with usual clinical diagnoses, since no differences were found in the prevalence of previous psychiatric diagnoses between profiles. Most patients in each profile received diagnoses in F4 (anxiety disorders) and F3 (mood disorders) ICD-10 categories, which were numerically the most common diagnoses in the sample. However, diagnoses of disorders with an onset during childhood and adolescence (eg, F8 and F9) and schizophrenia (F2) were overrepresented in profile 2. Profile 3 was particularly overrepresented among the small group of patients with organic mental disorders (F0) in the sample, which could implicate a more complex disease course. Interestingly, in a

previous paper, our medical team found that the assessments made by clinicians did not correlate well with patients' self-reports within 24 hours of a clinical evaluation [165].

The presence of sporadic suicide thoughts can be relatively frequent in psychiatric patients, but eHealth apps could help to identify profiles with higher suicide risk, such as profile 2. Previous literature has suggested improvements in mood, well-being, anxiety, and self-awareness, as well as a higher adherence to treatments among users of eHealth apps [135], [166]–[168]. Electronic assessment tools, such as the one used in our study, may support physicians to discriminate patients with high suicide risk in order to adjust their interventions.

Among the limitations of our study, we note the use of only baseline assessments and incomplete clinical information. The described profiles might not be reflective of eHealth users who continue to use the app regularly. Besides, our intention was not to map the participants onto Diagnostic and Statistical Manual of Mental Disorders (DSM) or ICD categories but rather to identify symptomatic profiles that are not necessarily reflected in psychiatric diagnoses. This study was designed to explore the utility of a new method to classify e-mental health users, and it needs to be completed with follow-up data. Nonetheless, once the SPFM is trained, it will be possible to analyze changes in patient profiles during continuous assessment with several time points. It will also be possible to link the electronic assessment with medical records. Our results could help to select the most performing questions according to mental disorders or patient profiles, which, in turn, could be used to create shorter and more efficient questionnaires. We can see in our study that the question about medication intake had very low informative value.

There are still many concerns regarding e-mental health that need to be addressed. One of the main concerns reported by both professionals and users is related to the privacy, ownership, and responsible use of medical information [169], [170]. This is one of the major challenges that eHealth needs to address by means of privacy-preserving technologies [171]. Accessibility and difficulties to find reliable sources of medical information are also important concerns in the population, especially among older adults [172]. Medical professionals also have doubts about the capacity of online information to improve the knowledge of patients and have reported concerns regarding the capacity of telemedicine to enhance physician-patient bond [170]. All these concerns must be addressed in order to improve the acceptability and use of eHealth tools. A recent study suggested that there is still a low preference for the use of eHealth tools among the adult general population [147]. However, those who have already used eHealth apps usually feel confident to continue using them. Some studies have reported a sense of security and the existence of a relational bond between eHealth apps and patients with psychiatric diseases [173], [174]. Our analyses show that machine learning can help to classify e-mental health users and provide clues for their diagnoses and, importantly, their needs in terms of treatment. If machine learning helps physicians to take clinical treatment decisions based on data, the social perception about available eHealth tools will certainly improve.

4.2. Disturbed Sleep as a Clinical Marker of wish to Die

4.2.1. Introduction

Suicidal behaviour is a major public health problem. Over 800,000 people take their own lives every year worldwide, and approximately 20 times more people attempt suicide [175]. Suicide risk assessment and monitoring are crucial for the prevention of suicidal behaviour [176]. However, a direct assessment of suicide risk can be difficult in contexts where a specialized clinical support cannot be guaranteed, such as in primary care settings, or when using self-report questionnaires [177], [178]. Clinical proxies of suicidal behaviour can facilitate assessment and be less overwhelming for patients. Wish to die —or passive suicidal ideation— has been shown to increase the risk for suicide attempts and death by suicide, even in the absence of active suicidal ideation [179]–[182]. A recent systematic review and meta-analysis deeply characterized passive suicide ideation, concluding that it was highly similar to active suicide ideation in terms of psychological correlates, and that it was strongly associated with suicide attempts [182].

There are some previous smartphone monitoring studies in suicide research, which have looked into different aspects such as fluctuation of suicide ideation over time or feasibility and acceptability of monitoring systems [183]–[191]. However, most of these previous studies have follow-up periods of less than a month and sample sizes of less than a hundred participants [184]–[189]. Another limitation is the use of economic incentives to increase engagement, a practice that limits the applicability of results [192], [193]. However, smartphone monitoring can be feasible in real-world conditions, as shown by a feasibility study by our research group: we tested the MEmind application in 457 participants and obtained a retention rate of 66.6% and an overall 68.0% compliance with questions over two months of observation, without using economic incentives [191].

Sleep is recently emerging as a promising clinical marker. Several forms of disturbed sleep —including insomnia, nightmares, poor sleep quality and reduced sleep quantity— have been associated with several forms of suicidal thoughts and behaviours (STB)—including wish to die, suicidal ideation, suicide attempts and death by suicide [194]–[198]. This association has been confirmed in a number of systematic reviews and meta-analyses, which have found a significant and independent association between sleep disturbances and STB [199]–[202].

The association between sleep problems and STB seems relevant in both the long term and the short term. For example, a 13-year retrospective study of 479,967 patients showed that insomnia tripled the risk of suicide attempt [203]. Looking at the short-term relationship, a prospective study with a 21-day follow-up showed that actigraphy-measured sleep variability was a significant predictor of suicidal ideation [196]. Another study, where participants completed a sleep diary over the course of 1 week showed that short sleep duration and poor sleep quality increased the severity of next day suicide ideation [198]. Thus, sleep may increase the risk of suicide both over several years and over the course of

just a few hours. The long-term association may be mediated by increased risk for other mental disorders, such as depressive disorders, among other factors, while the short-term association may result from emotional dysregulation and impulsivity [201].

Negative feelings and appetite alterations have also been associated with STB in previous studies. Negative feelings have been explored in some previous smartphone monitoring studies [184], [186]–[188], [204]–[206]. For instance, in the study by [187], hopelessness and perceived burdensomeness were prospectively associated with suicide ideation. However, this is not the case for appetite, which has been explored with more traditional methodologies —for instance, [207] found that adolescents with appetite loss had a five-fold risk for suicidal ideation compared to those with normal appetite— but has not yet been measured through smartphone monitoring.

Smartphone monitoring may be particularly suitable for measuring the relationship between sleep and suicide, given the huge variability of suicidal ideation over short periods of time and the probable role of sleep as a short-term marker of STB.

As shown in the meta-analysis by [208], there has been little progress in the search for valid risk factors for suicidal behavior in the last decades. The authors of this meta-analysis highlight the potentials of machine learning to advance in the field of suicidology [208].

4.2.2. Objectives

In this study, we use the same model as in the previous section but another smartphone questionnaire to explore the associations between wish to die, disturbed sleep, negative feelings and altered appetite, in non-incentivized psychiatric patients over three months of observation. Our hypotheses are:

1. That we will be able to detect relevant latent features across the dataset as we did before

2. That these latent features will show a prominent association between sleep problems and STB (i.e. these variables will be present in a short time window).

4.2.3. Data & settings

This is a prospective cohort study of psychiatric patients receiving mental health care at the *Hospital Universitario Fundación Jiménez Díaz* as in the previous work (Section 4.1). This study comprises a sub-set of the cross-national multicentre study SmartCrisis. SmartCrisis' study protocol has been published elsewhere [136].

The study was approved by *Ethics Committee* of the Jiménez Díaz Foundation University Hospital and was conducted according to the principles set forth in the Declaration of Helsinki [209]. All patients gave written informed consent to participate.

Sample

Participants were outpatients with any psychiatric diagnosis who were approached during regular appointments with their psychiatrist or psychologist and invited to participate. Inclusion criteria were being aged 18 year or older, having a history of suicide behaviour and/or suicidal ideation measured with the Columbia Suicide Severity Rating Scale (CSSRS) [210], being able to understand and sign the informed consent form, owning a smartphone with internet access.

Again, the questionnaire was administered through a smartphone application: the **MEmind Wellness Tracker** (See Section 4.1.2 for details)). This time, the questionnaire is called Ecological Momentary Assessment (EMA) and has 32 items, which were asked with different frequencies and in different order. Suicide-related variables were 'Wish to die', and 'No wish to live'. Including more direct questions about suicide intentionality was considered during the design of the study, but this option was rejected by the Ethics Committee. Three other areas were explored: negative feelings (13 questions), sleep quality and quantity (10 questions), and appetite (7 questions). Questions about wish to die, wish to live and negative feelings were based on the Salzburg Suicide Questionnaire [211]. Questions about sleep were extracted from the Insomnia Severity Index (ISI) [212], and questions about appetite were extracted from the Council on Nutrition Appetite Questionnaire (CNAQ) [213]. In order to increase engagement without using incentives, we did not ask the same questions every day. Instead, we rotated questions, making them different from day to day. The questions were randomized, with the guarantee that all were asked throughout the follow-up a certain number of times (this frequency changes according to the questions, as shown in Figure 4.7), but the order in which they will be asked varies from patient to patient. We did this to lessen the burden for the patient and thus allow for a longer follow-up period. During the first month, the evaluation included 4 questions. Afterwards, evaluation included 2 questions. Every day at 10 am there was a question about sleep regarding the previous night. The rest of the questions were asked at random times from 10 am to 10 pm. They all appeared as a notification on the users' screen. Reducing the burden of questions is a fundamental aspect in a smartphone monitoring study with a longer than usual follow-up period. The pre-processing and algorithm employed allowed us to compensate for missing data resulting from this approach.

Questions about sleep and wish to die/live were over-represented according to our hypotheses. The variables explored and the frequency of administration of questions are shown in Figure 4.7. The complete questionnaire is shown in Appendix C.

Other measures

During the baseline interview, a trained psychologist assessed suicidality, sleep disturbances, mood and anxiety symptoms using the following standardized questionnaires: CSSRS [210], suicidality module of the Mini International Neuropsychiatric Interview

1	Psychological pain		17	Difficulties staying asleep	**
2	Stress		18	Early morning awakening	**
3	Restlessness		19	Noticeability of sleep difficulties by others	
4	Hopelessness		20	Negative feelings when awakening	**
5	Self-hatred		21	Poor sleep quality	**
6	Hatred		22	Overall sleep dissatisfaction	
7	No wish to live	*	23	Preoccupation and distress caused by sleep difficulties	
8	Wish to die	**	24	Interference of sleep problems with daytime functioning	
9	Wish to have a close person whom to talk to		25	Daily fatigue caused by sleep difficulties	
10	Thwarted belongingness		26	Decreased appetite	**
11	Impression that other make decisions for you		27	Fullness after eating	
12	Need of affection		28	Increased appetite	
13	Inability to help family		29	Tastelessness of food	**
14	Inability to help others		30	Distortion of food taste	
15	Emotional disconnection		31	Increased number of meals	
16	Difficulties falling asleep	**	32	Nausea after eating	

First month	
Number of questions per day	4
Questions asked at least twice per week	*
Questions asked at least once every four days	**
Questions asked at least once every two weeks	Rest

Afterwards	
Number of questions per day	2
Questions asked at least once per week	*
Questions asked at least once every eight days	**
Questions asked at least once every six weeks	Rest

Figure 4.7: Variables assessed through smartphone monitoring and assessment frequency.

7.0.0 [214], Pittsburgh Sleep Quality Index (ICSP) [215], ISI [212], Inventory of Depressive Symptomatology (IDS) [216], Young Mania Rating Scale [217], and State-Trait Anxiety Inventory (STAI) [218].

Sociodemographic variables, including age, sex, marital status and employment status, were also collected. Diagnosis was established clinically, based on information collected in the electronic medical record, which in turn was based on the tenth edition of the ICD-10 criteria. The diagnoses were catalogued into ten diagnostic groups as per the ICD-10.

Procedure

Recruitment took place from February 2018 to January 2019. Data collection was performed by a team of research psychiatrists and psychologists different from the clinicians that cared for the patients. During the baseline visit, participants had the application installed in their mobile phones and were taught how to use it. Patients were followed for a median of 89.8 days, resulting in $9,878$ person-days.

4.2.4. Methods

We applied the non-parametric Bayesian method, SPFM, that we used in Section 4.1 to analyse the data. In this scenario, this algorithm detects sets of variables that tend to adopt certain altered values in the same time frame (in this case, after testing the model, the most appropriate time frame was established at 96 hours). Again, here, we call these sets of

variables “latent features”. Thus, features are supravariables formed by the grouping of variables. For instance, if two or more given variables are frequently present at the same time, the method would label them as a relevant latent feature. Based on the presence or absence of these latent features over time, the system takes the next step of clustering, describing “profiles”, which are the sum of latent features.

One way to understand this concept is to compare it with the Myers- Briggs Type Indicator (MBTI) personality test. The MBTI is composed of 64 questions (whereas our questionnaire is composed of 32 questions) and depending on the answers, 8 personality ‘traits’ are described (Introverted/Extroverted, Sensitive/Intuitive, etc.). These ‘traits’ would be the equivalent to our ‘features’. From the combination of ‘traits’, 16 personality ‘types’ emerge (ISTJ, ISFJ, etc.). These personality ‘types’ would be the equivalent to our ‘profiles’.

As in the MBTI, features and profiles are individual. The difference is that in our case the features and profiles are not predetermined theoretically, but rather are empirically determined through the model after observing the dataset. The number of features to be detected is also not predetermined: the algorithm detects them for each data set. Another difference is that, while personality is (relatively) stable, both features and profiles are dynamic: each person can have different features and profiles depending on the socio-familial and intrapsychic context they are going through at any given time (and the mechanisms that trigger a change in profile may be different for each person), although profiles are more stable than features.

As we know, the SPFM discovers these latent features through a sparse analysis. That is, only a small subset of data will offer discriminant information. Using this method we can surpass the limitations caused by the turn-over system of the questions, including missing data, sparsity of valuable information, and chronological irregularity of the assessment.

Data processing

Since the SPFM method uses dichotomous variables, once again, all questions were transformed. In addition, in this work we had different kind of questions from different questionnaires, so the format of the answers could be from continuous values until categorical ones.

First, the value of some questions were reversed so the highest punctuation means the worst state of the user who answers. These modified questions are marked in Appendix C.

Second, we have dealt with the temporal component by grouping data from questions performed during some consecutive days. For that, we have tried windows from two to four days, so the model will consider that everything answered in that period of time belongs to the same sample. For the results presented in this work, the window was set to four days.

Third, all questions have been normalized to continuous values between 0 and 1 and then transformed them to binary values. In this work, we established a threshold of 0.5, so

that users with a score equal to or greater than 0.5 would obtain a value of 1, equivalent to showing the worst possible status for the particular question. The rest were set to 0.

To address the issue of multiple testing, we estimated the False Discovery Rate using the Benjamini and Hochberg procedure [219]. To correct the overrepresentation of some of the questions, adjustments were made by histogram equalization and subsequent remodeling.

4.2.5. Results

Characteristics of the sample

Of the 189 patients approached, 165 (87.3%) agreed to participate in the study. The mobile application could not be installed in 26 (13.8%) participants due to technical issues. 139 (73.5%) had the application installed. Participants who answered for less than five days were excluded, resulting in 110 participants included in the final analysis. There were no significant differences regarding age, sex, or history of STB between compliant and non-compliant participants. Total number of responses across the 110 participants was 13,959. Number of responses ranged between 6 and 657. Mean number of responses was 126.9 (standard error = 11.44). Estimated response rate was 52.88%.

The mean age of the participants was 45 years. Gender distribution was 65.6% female and 34.4% male. The most frequent diagnoses were anxiety disorders (64.8%), mood disorders (51.2%), and personality disorders (51.2%). 92% of patients reported previous suicide attempts —54.4% had attempted suicide one or two times, and 36.8% three or more times—, while 8.0% had only suicidal ideation and no previous attempts. Table 4.5 shows the full description of the sample.

After the preprocessing of the data, with the windowing, we kept 2428 observations belonging to all patients.

	n (total=165)	%	Mean (SD)
Gender			
Male	57	34.4	
Female	108	65.6	
Age (years)			44.3 (15.22)
Marital status			
Married/Coupled	65	39.2	
Single	68	41.5	
Separated/Divorced	30	18.1	
Widowed	2	1.2	
Employment status			
Employed/Student	72	43.7	
Unemployed	37	22.2	
Retired	13	7.8	
Temporal leave	28	17.3	
Permanent leave	15	9.0	
ICD-10 Psychiatric diagnosis			
Mental disorders due to drug use	21	13.0	
Psychotic disorders	1	0.6	
Mood disorders	83	51.2	
Anxiety disorders	105	64.8	
Personality disorders	83	51.2	
Other	29	18.0	
Number of previous suicide attempts			
None (only previous suicide ideation)	10	8.0	
1-2	68	54.4	
3 or more	46	36.8	
CSSRS (SI subscale) lifetime			4.33 (0.92)
CSSRS (SI subscale) last month			2.63 (1.96)
IDS			23.82 (11.96)
ISI			13.13 (6.65)

Table 4.5: Baseline characteristics of the sample.

Latent features analysis

We identified four relevant latent features in the dataset—i.e.: groups of variables with high probability of scoring positive within the same time window (96 hours)—. Following the simile proposed in the methodology, these would be the "personality trait" found. Figure

4.8 shows the four latent features and the high-scoring variables that characterize them. As we know from the previous work, each feature is made up of a score on each of the variables. This score is the result of dichotomizing the variables and it is related with the probability of scoring more than 0.5 on the continuous variables. Each of these features is explained in detail below:

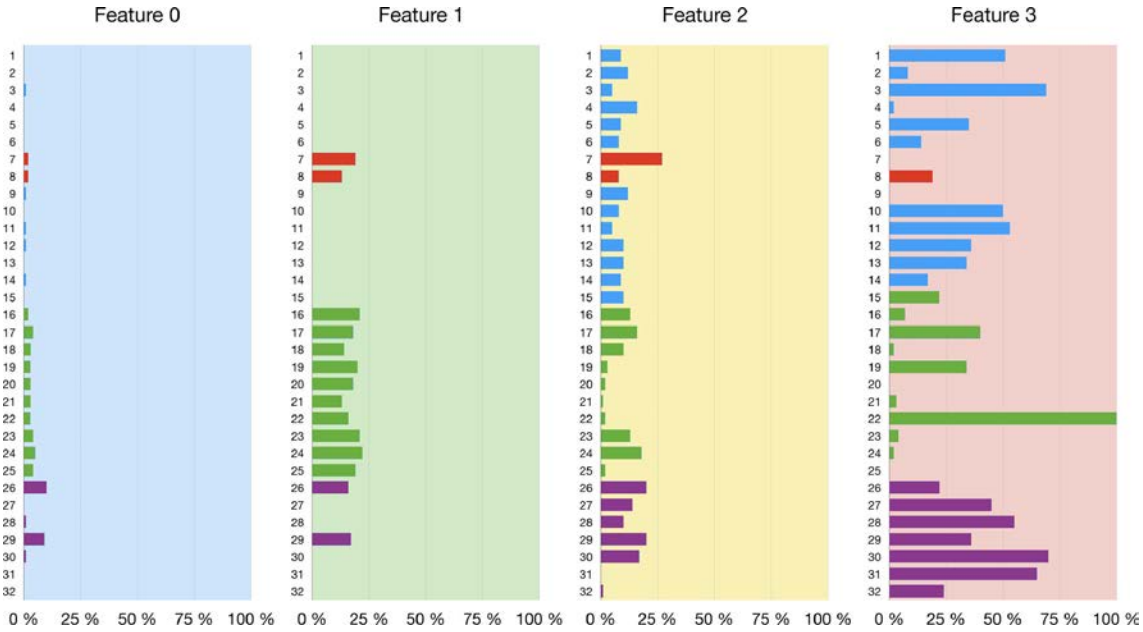


Figure 4.8: Latent features (0, 1, 2 and 3), characterized by different scores on each of the variables.

- Feature 0 represents the basal or resting state. This is a state common for all the dataset (i.e., these are the fixed scores of adopting each of the variables throughout the follow-up) and as in the previous work, it is called bias. Then for this feature 0, we observed low scores in all the variables. The highest scoring variables were 'Decreased appetite' (26), 'Tastelessness of food' (29) and 'Interference of sleep problems with daytime functioning' (24).

- Feature 1 is characterized by high probability of scoring positive in the variables 'No wish to live' (7) and 'Wish to die' (8), along with moderately increased probability scores in all sleep-related variables. 'Decreased appetite' (26) and 'Tastelessness of food' (29) also presented an increased probability score.

- Feature 2 is characterized by moderately increased scores in most variables, indicating worse general state. 'No wish to live' (7) had a 27% probability of scoring positive. 'Wish to die' (8) also presented an increased probability score, as did sleep-related variables such as 'Difficulties falling asleep' (16) or 'Difficulties staying asleep' (17), negative feelings such as 'Hopelessness' (4), and appetite-related variables such as 'Decreased appetite' (26).

- Feature 3 shows high probability scores in most variables to a greater extent than feature 2. The highest scoring variables were 'Overall sleep dissatisfaction' (22) (100% probability of scoring positive), 'Restlessness' (3) and 'Distortion of food taste' (30).

Based on these relevant latent features, we observe that in those features where 'Wish to die' is increased, sleep problems are also increased. That is, 'Wish to die' and sleep problems have an increased probability of adopting positive values in a time window of 96 hours. As mentioned in the Section 4.2.4, the next step is to determine the most relevant profiles based on the sum of the different features throughout the monitoring period. Thus, the profiles are the combinations of features —equivalent to the sum of the scores of each of the features— that are more usual over the follow-up period. The clustering method over the SPFM result revealed five relevant profiles. The profiles are described in Table 4.6. The feature 0 is present in all profiles, as this is the basal state of the sample. Profile 1 involves only feature 0, and it was the most frequent one throughout the follow-up. The profile 2 is characterized by presenting Feature 0 plus Feature 1. It was the second in frequency. Profile 3 (third in frequency) is characterized by the presence of Feature 0 plus Feature 2. Profile 4 (4th in frequency) comprises three latent features (0+1+2), and profile 5 (the rarest) comprises the sum of all four latent features at the same time.

Profile	Number of observations	Feature 0	Feature 1	Feature 2	Feature 3
1	1452	Yes	No	No	No
2	738	Yes	Yes	No	No
3	142	Yes	No	Yes	No
4	93	Yes	Yes	Yes	No
5	3	Yes	Yes	Yes	Yes

Table 4.6: Profiles identified in the sample and features present in each profile.

Although profiles are more stable than features, each participant can evolve from one profile to another throughout the follow-up. Figure 4.9 shows this variability. Thus, each of our 110 patients is represented on the x-axis; the y-axis represents the profiles and the color indicates the probability of adopting each of these profiles throughout the follow-up.

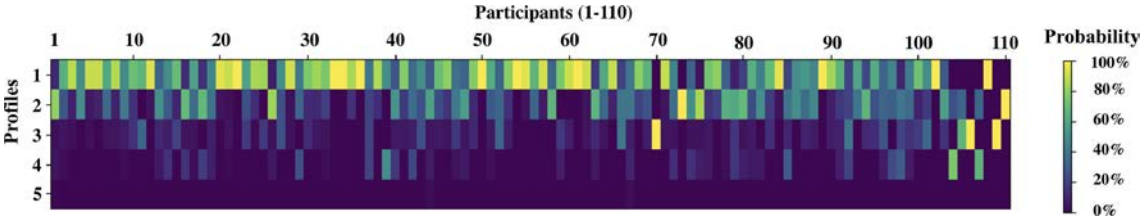


Figure 4.9: Heat map of the evolution of profiles. For each patient (horizontal axis) the figure shows the probability (colour legend) of presenting each of the five profiles (vertical axis) throughout the follow-up period.

4.2.6. Discussion

Comparison with previous findings

In our study, we found that 'wish to die' and sleep problems tend to be present at the same time window of 96 hours. Our results add to the evidence about the association between sleep disturbances and different forms of suicidal behavior [199]–[201]. Particularly, our study adds to the evidence that sleep problems are associated with STB in the short term [196], [198]. Although in this work we have not studied the directionality of the association, we have observed that both variables tend to occur together within 96 hours. In a sleep diary study performed by [198] in 51 patients, authors found an unidirectional association between short sleep duration and suicidal ideation within 24 hours. In contrast, suicidal ideation did not predict sleep problems [198].

Lack of sleep leads to emotional dysregulation [220], [221], which may partially explain its association with passive suicide ideation Additionally, insufficient sleep increases impulsivity [222], [223]. If we consider the integrated motivational-volitional of suicidal behavior [224], we see how impulsivity is one of the main precipitating factors of suicidal behavior. Sleep problems, via increased impulsivity, may act as precipitating factors for suicidal behavior. Therefore, the joint presence of suicidal ideation and sleep problems, with the impulsivity that they entail, could be indicating a risk of progressing from the motivational phase to the volitional phase. One line of research that we intend to address in future studies is to explore whether data recorded through smartphone monitoring correlate with clinical events, such as suicide attempts. In this way we aim to further characterize sleep problems as a risk marker for suicidal behavior.

Other factors associated with wish to die

Low appetite was also associated with high wish to die and low wish to live, specifically the components of "No appetite" and "Bad food taste". Previous studies have found an association between alterations in appetite and suicidal behavior [207]. As for the questions about negative feelings, they scored high in profiles where the remaining symptoms –including wish to die– were also present, suggesting a non-specific pattern of clinical severity. Negative feelings have been associated with suicidality before, especially those related to perceived burdensomeness, hopelessness, and thwarted belongingness, both in the short-term [204] and in the mid-term [225]. Thus, in smartphone monitoring studies with short follow-up periods (1–2 weeks), negative feelings appeared concurrently and/or prospectively associated with suicidal ideation [204], [206], while in the study by [225] perception of entrapment predicted suicide ideation in the course of 7 weeks. While the evidence from this and other studies suggests the potential of sleep problems as short-term predictors of suicidal behavior, negative feelings may be related to suicide more broadly along the suicidal trajectory. In the integrated motivational-volitional model of suicidal behaviour [224], negative feelings, such as entrapment, burdensomeness or thwarted

belonginess, act in the intermediate, motivational, phase. Sleep problems and negative feelings could also interact with each other. For instance, a recent study showed that healthy sleep decreased the impact of psychological distress on suicide risk [226].

The advantages and disadvantages of smartphone monitoring

Despite its numerous advantages, smartphone active monitoring also presents with limitations. An intrinsic weakness of this methodology is participant fatigue [227]. Submitting them to the same questions every day causes patients to abandon the use of the application. In the contrary of the previous work where all questions were answered each time, in this study, we asked a set of questions on each shot. Furthermore, we tackled the issue by incorporating a turnover system to avoid repetition of questions and thus improve engagement. To be able to analyze the data obtained in this way and compensate for the resulting missing data, we used a IBP based model which has been scarcely used before. The specific variation we introduced —Sparse Poisson Factorization Model— is novel in psychiatric research and this is our second work where we apply it.

Another way to decrease fatigue is to resort to smartphone passive monitoring, using the smartphones' native sensors, without the active collaboration of the user. This methodology has been explored before in suicide research [183], [191], as well as in other mental health areas [228], [229]. However, there are still issues to resolve, such as the validation of the sensors for measuring variables such as mobility or sleep, and the usefulness of these variables as clinical proxies of STB. In addition to its usefulness in research, smartphone monitoring technology could be harnessed for clinical practice through the development of a clinical monitor that alerts clinicians of imminent risk. However, there is still a long way to go before they can be fully implemented in clinical practice.

Strenghts

To our knowledge, this is the first smartphone monitoring study exploring the association between passive suicide ideation and sleep problems —two previous smartphone-based studies included sleep related questions in their protocols [188], [191], but both of them were feasibility studies and did not offer results regarding the association—. This study is also innovative because it incorporates a rotation system in the questions that reduces repetition. Other studies also rotate their questions; for example, [230] conducted 3 assessments per day, each of them including 3 randomly chosen questions from the Patient Health Questionnaire-9. However, in their case, each day they ended up asking the full nine questions of the questionnaire. In our case, the patients were only asked 2 to 4 questions at random from a pool of 32 questions. This considerably reduces the burden at the cost of generating missing data. These missing data can be compensated for by using the suitable model as our case. Our study has been performed in real-world settings, with larger sample sizes and longer follow-up periods than previous studies.

Limitations

Our findings must be considered in light of some limitations. Sleep assessment was based on participants' perception and thus subjective. We did not ask about suicidal ideation directly, but instead used the proxy "wish to die", following the recommendation of the Ethics Committee. This was deemed as less overwhelming for patients, taking into account that ours was a continuous, digitally delivered assessment, without direct clinical supervision. Another limitation is that the SPFM compensates for the missing data at the cost of converting continuous variables in dichotomous variables, thus losing variance. Also, the associations were not explored longitudinally, but cross-sectionally within a short reference period (variables that appeared together within the same 96 hours). Thus, we cannot establish directionality of the association. Finally, the study was conducted in patients at high risk by virtue of prior suicide attempts or suicidal ideation. Studies including larger samples from more general psychiatric or other populations would be instructive.

4.2.7. Conclusions

Using smartphone monitoring and machine learning techniques we found that disturbed sleep was associated with wish to die among psychiatric patients in the short-term (within 96 hours). Our findings stress the importance of evaluating sleep as part of the screening for suicidal behavior. In the presence of other factors, such as high-risk psychiatric diagnoses or a history of previous attempts, it is relevant to ask about the quality and quantity of sleep, as it can be a precipitating factor. This is relevant not only for mental healthcare, but also for primary-care settings considering that nearly half of patients who attempt suicide visit their general practitioner in the previous month [231]. In contrast, attempters rarely make a specific comment about their suicidal thoughts during these visits [232]. Questions about sleep may be less threatening to the patient and easier for non-specialized physicians to address.

The refinement of statistical methods and the advancement of technology can increase the potentials of smartphone monitoring, including the future development of smartphone-delivered clinical interventions. Statisticians are increasingly calling for a research landscape beyond the p value [233]. Machine learning techniques represent an advance from traditional statistical methods.

This line of research can result in innovative preventive and therapeutic tools that make use of new technologies for the detection and prevention of suicide. However, we must acknowledge the need for more work before these tools can be applied in everyday clinical practice. One of the main challenges is the integration of these alternatives within broader treatment plans, taking into account factors such as digital literacy or the time of the clinical course to select which patients that can be benefit the most from these technologies.

4.3. Suicidal High Risk Patients State during COVID-19

4.3.1. Introduction

This chapter corresponds to the last work in my line of studies of applying a probabilistic method to smartphone-based health questionnaires to psychiatric patients. It emerged during the COVID-19 quarantine and consisted in applying the same method as the previous two chapters, the SPFM, to find the profiles distribution of patients before and during the lockdown.

Psychiatric patients are particularly vulnerable to the psychological impact of the coronavirus disease 2019 (COVID-19) outbreak. Social distancing and lockdown measures result in multiple stressors known to increase risk for suicide, including social isolation, financial stress, decreased access to mental healthcare and medical comorbidities[234]. Research on the mental health consequences of this crisis is considered a priority[235]. However, quarantine has interfered with face-to-face research. Mobile technology applied to health – known as mobile health or m-Health – can overcome these barriers. In this study we use smartphone-based EMA to explore the impact of COVID- 19 social distancing and lockdown measures on suicide risk, in a sample of psychiatric patients at high risk for suicide.

4.3.2. Method

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. All procedures involving human patients were approved by the Ethics Committee of the University Hospital *Fundación Jiménez Díaz*. All participants provided written informed consent to participate in the study.

Participants and procedures

Using EMA, we prospectively assessed 36 adult patients, who were being treated at our suicide prevention out-patient clinic because of a high risk of suicide. EMA was delivered using the MEmind smartphone app, which is available for both Android and iOS operating systems. As in the previous work, EMA questions were announced as push notifications on users' screens. A detailed description of the MEmind app has been published elsewhere[136], [158]. Participants were recruited from an ongoing multisite study examining longitudinal risk factors for suicide (SmartCrisis[136]).

Patients were included in the study if they had a history of at least one suicide attempt or an emergency department visit because of suicidal ideation. Written informed consent was obtained from all patients. Pseudonymization of the participants' personal data was

employed, by using a unique identification code for each participant. The follow-up period was divided into: (a) pre-lockdown: 1st October 2019 to 13 March 2020 (before the implementation of Covid-19 lockdown measures); and (b) lockdown: 14th March to 14th April 2020.

At baseline and at follow-up, patients were administered the CSSRS[210]. To safeguard the well-being of our patients, upon detecting an alarming level of suicidal ideation (threshold was established at CSSRS suicidal ideation subscale score ≥ 4), their attending psychiatrist was informed, and it was suggested to patients that they attend the emergency department.

The EMA questionnaire

During this work we use the same questionnaire as in the previous part so refer to Section 4.2.3 for more details. Appendix C shows all the questions and their scoring. The MEmind EMA questionnaire has shown good acceptability in preliminary studies[191], [236]. As constant repetition of questions can place a significant burden on the user, we have incorporated the same turn-over system for questions as we used in previous work. Out of the pool of 32 questions participants were asked two to four random questions every day, at random times from 10.00 to 22.00 h. Figure 4.10 shows the variables explored in the EMA questionnaire and the frequency with which the questions were asked.

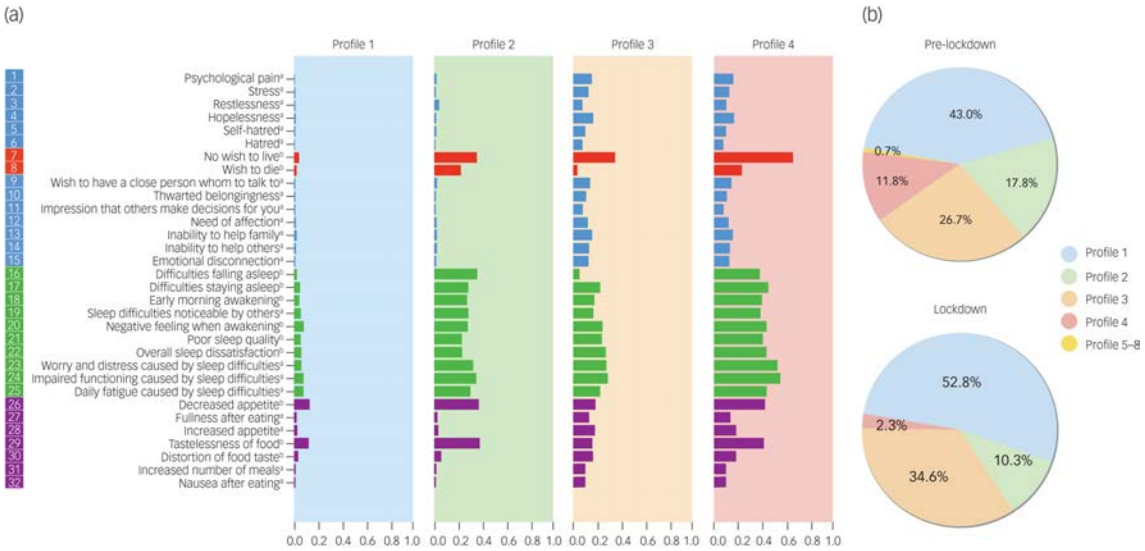


Figure 4.10: (a) Suicide risk features identified using the SPFM. Vertical axis: variables. Horizontal axis: probability of scoring positive each of the variables. (b) Distribution of features before and during lockdown.

Assessment frequency: [a] at least once every two weeks during the first month and at least once every six weeks afterwards; [b] at least twice per week during the first month and at least once per week afterwards

Statistical analysis

We used the same SPFM that we did in Sections 4.1 and 4.2. As before, the features revealed by the SPFM are supravariables formed by the grouping of variables, that is: sets of variables that tend to adopt certain abnormal values in the same time frame. They are dynamic, and the same person may have different features over time. So, one of the main differences regarding previous works is the time treatment of the samples. In this work, we separate patients data in two sets, before and during the lockdown, and compare the profiles found in both periods. More precisely, we compared individual suicide risk features before and during the lockdown.

The values of each of the 32 questions were standardised so that all were again expressed as 0 or 1 and the highest value would always express a worse state of mental health. The negative connotation of the profiles is present in the three works from this line.

4.3.3. Results

Sample

Mean age of the participants was 41.7 years (s.e. = 16.3). The majority of the participants were women (n = 31; 86.1%). The most common psychiatric diagnosis was mood disorders (n = 21; 58.3%). The mean number of previous suicide attempts was 1.1 (s.e. = 0.2)

Profiles

After applying the SPFM we obtained 4 different features, the bias and three more. At the contrary that in the two previous works, after doing the clustering, we got a quantity of 8 profiles with different combination of features active. However, we identified four profiles that accounted for more than 99.5% of the participants' responses (see Figure 4.10(a)), so we focused our results in those. All of them have the bias active, profile 2 also has feature 1, profile 3 has feature 2 and profile 4 has both features 1 and 2 active. The forth feature is only present in profiles 5-8 so it lacks of interest for us.

Regarding the profiles definition, we found that profile 1 is characterised by low values (i.e. low probability of scoring positive) across all 32 suicide risk factors. Profiles 2 and 4 are characterised by a high desire for death, lack of wish to live, decreased appetite and tastelessness of food, and sleep problems; Profile 4 also shows high values for negative emotions. Profile 3 is characterised by lower desire for death, and lower appetite and sleep symptoms, with high values of negative emotions.

Before quarantine, the most prevalent profile was the number 1, with 43.0%. That is, of the 960 responses before quarantine, 43.0% were grouped in profile 1. The second most prevalent was profile 3 (26.7%), followed by profile 2 (17.8%). Profile 4 accounted for 11.8% of the responses and the remaining profiles (5 to 8) accounted for 0.7% (95% CI

0.3–1.5%) of the responses.

During the quarantine, the dominant profile continued to be the 1 (52.8%). That is, of the 214 responses, 52.8% were grouped around profile 1. This represents a 22.8% increase. The second most common profile was still profile 3, with 34.6%. This represents a 29.6% increase. Profile 2 fell to 10.3%, a 42.1% decrease, and profile 4 fell to 2.3%, an 80.1% decrease. The remaining profiles (5–8) were not represented during the quarantine.

Results of the χ^2-test show there are statistically significant differences before ν. during lockdown (Profile 1: $\chi^2 = 6.38$, P = 0.012; Profile 2: $\chi^2 = 6.69$, P = 0.010; Profile 3: $\chi^2 = 5.04$, P = 0.025; Profile 4: $\chi^2 = 16.20$, P < 0.001).

4.3.4. Discussion

Contrary to our expectations, we observed that self-reported suicide risk appeared to decrease during a COVID-19-related lockdown period, in a prospective cohort monitored using smartphone-delivered EMA. Specifically, we found a decrease in the wish to die, and in the rates of appetite and sleep symptoms.

Strengths and limitations

Strengths of our study include the prospective design and real-time monitoring of dynamic suicide risk using EMA. Our results should be interpreted with caution given the modest sample size. This modest sample size may be the reason why we have found an uneven gender distribution, with over 85% of patients being women. However, in a prior EMA study by our research group they also found a predominance of women in the sample[191]. Another potential limitation, in the same way that in the work from Section 4.2, is that we did not ask directly about suicide intent but employed the indirect measure 'wish to die'. However, a recent systematic review and meta-analysis exploring passive suicide ideation found that it was highly similar to active suicide ideation and that it was strongly associated with suicide attempts[182]. Also, the observation period before lockdown was longer than during lockdown. Finally, the length of the follow-up period was not uniform across the sample.

Comparison with findings from other studies

Other studies have also found a decrease in suicidal ideation as a result of COVID-19-related measures. For instance, a recent study showed that internet search queries related to suicide decreased after the USA issued stay-at home-orders[237]. Although it may seem surprising that suicidal ideation decreases, it is actually consistent with some previous studies showing a drop in suicide rates during periods of social emergency, such as wartime or terrorist attacks[238], [239]. However, there is also evidence indicating that this decrease may be just temporary: the study by [240] shows that, although there is a decrease in

suicidal behaviour during wartime, just after wars end, suicidal behaviour increases to levels higher than those observed before the war. Thus, during the post-war period, the harmful effects of conflict on an individual's mental health become apparent. In the same way, the possibility exists that there will be an increase in suicidal ideation and behaviour above the expected level once the acute COVID-19 crisis ceases. We must be prepared for this contingency.

Implications

Continuity of care has been affected by the COVID-19 crisis. In order to minimise the risk of contagion, non-urgent face-to-face consultations have been discontinued in many countries, including Spain. Telemedicine allows us to continue to provide mental healthcare services to our patients. New technologies are already being used to preserve people's mental healthcare during the COVID-19 crisis, for example in the form of online services[241].

Ensuring access to adequate mental healthcare for vulnerable populations, such as psychiatric patients at high risk for suicide, should remain a priority during times of social emergencies. Smartphone-based monitoring can be used to monitor high-risk populations during social distancing and lockdown periods.

5. ROBUST SAMPLING

5.1. Introduction

Deep learning requires regularization mechanisms to reduce overfitting and improve generalization. We address this problem by a new regularization method based on distributional robust optimization. The key idea is to modify the contribution from each sample for tightening the empirical risk bound. During the stochastic training, the selection of samples is done according to their accuracy in such a way that the worst performed samples are the ones that contribute the most in the optimization. We study different scenarios and show how it can make the convergence faster or increase the accuracy.

Machine learning algorithms assumed that the samples are coming iid (independent and identically distributed) from $p(\mathbf{x}, y)$ and hence they use the samples equally during training. For example, in deep learning all the samples enter with the same probability in each of the mini-batches [2]. But not all samples are equally relevant when learning classifiers and regressors because some of them might be hard or easy to classify or they might be under-sampled or over-sampled in the training set without our knowledge. There are many ways in which non-uniform sampling can be used to improve convergence speed or quality by relying on non-uniform sample. The first example that comes to mind is AdaBoost [242], which uses different weights for each training example to build a robust classifier.

More recently there has been proposals to use importance sampling for training classifiers and regressors to reduce the variance of their estimates. In a nutshell, the objective is to increase the number of times a hard-to-learn sample appears in the mini-batch so the learning algorithm can converge faster and then weight their error by the number of times it has been used. For example in [243], the authors developed a non-uniform importance sampling technique to solve an online optimization problem with bandit feedback. In [12] the authors used the data structure to adapt the gradients of each observation. And [244] try to sample the datapoints from a non-uniform distribution according to a multiarmed bandit framework.

In the recent award-winning [245], the authors proposed bounds to reduce the variance of classifiers by relying on non-uniform sampling of the training set, but they did not compensate for the over-sampling (or under-sampling) of the training set in their bound. The non-uniform sampling is a feature that should make the learned classifier more robust and reduce the variance of its prediction. The results in [245] are mainly theoretical and they illustrate their algorithm in an example with very few training samples, large input dimension and using a logistic-regression classifier.

In this work, we embark on an implementation of this algorithm for training deep

learning models to understand if this theoretical result shows significant improvement for standard deep learning classifiers. We first propose two different alternatives on how to incorporate the non-uniform sampling within the mini-batches used in deep learning, leading to different ways in which hard-to-classify examples are repeated in the mini-batches. We then compare these algorithms with a standard optimization of neural networks. We have relied on well-known architectures and datasets not to biased our results with new neural networks or data. We found that there are some minor improvements in the convergence speed and reduction of error. However, this improvements seem to be more relevant in scenarios where the number of training samples is low. On the other hand, there is not a consistent setting for the hyper-parameters in our algorithms, what make it costly sometimes to find the best configuration. The proposed algorithms do not seem to hurt either the baseline performance and their computational complexity is negligible compared to the training of the neural network.

Also, we have noticed that if we do not use dropout [57], the improvement from using non-uniform sampling is significant. The improvements gains provided by dropout are equivalent to those of using our proposed implementation for [245]. Even though, both methods are thought for reducing the variance of the learnt models, they achieve comparable results by the completely different means.

[246] proposed a similar application, unifying importance sampling and minibatching algorithms so to assign some probability distributions to the samples of a set of minibatches and sample them. They propose a sampling scheme to improve the converge rates but, unlike us, they use probabilities to sample more relevant examples.

In a Bayesian setting, non-uniform sampling has been proposed in [247]. In it, the authors took a probabilistic approach in order to make inference by raising the likelihood of each data point to a weight. But in this paper the authors assume that the hard-to-learn samples are outliers that would contaminate the solution of the classifier and the algorithm actually under-samples them. The goal of sampling in this case is to reduce the outliers and not to make the classifier more robust to hard-to-classify examples that are still valid samples.

The rest of the work is outlined as follows. We review the main results in [245] in Section 5.2 and the proposed algorithms are detailed in Section 5.3. We then present extensive empirical results in Section 5.4. We conclude the paper in Section 5.5.

5.2. Motivation

5.2.1. Variance-based robust regularization

[245] proposed an alternative to empirical risk minimization that provides a robust and computationally efficient solution for small data sets. Particularly, it is based on tightening

the empirical risk bound by adding a variance term in the form

$$\frac{1}{n}\sum_{i=1}^{n} l(\theta, x_i) + C\sqrt{\frac{2\rho}{n}\mathrm{Var}_{\hat{P}_n}(l(\theta, X)}$$

where l is loss function, C is a parameter that depends on l and the desired confidence guarantee, and $\mathrm{Var}_{\hat{P}_n}$ the empirical variance.

5.2.2. The empirical risk extension

Instead minimizing this regularized risk functional, generally not-convex, the authors define a robust regularized risk

$$R_n(\theta, \mathcal{P}_n) = \sup_{P \in \mathcal{P}_n} \left\{ \mathbb{E}_P[l(\theta, X)] : D_\phi(P \| \hat{P}_n) \leq \frac{\rho}{n} \right\}$$

where D_ϕ is the ϕ-divergence with $\phi(t) = 1/2(t-1)^2$. The robust regularized risk is shown to be equivalent to

$$R_n(\theta, \mathcal{P}_n) = \mathbb{E}_{\hat{P}_n}[l(\theta, X)] + \sqrt{\frac{2\rho}{n}\mathrm{Var}_{\hat{P}_n}(l(\theta, X))} + \varepsilon_n(\theta) \tag{5.1}$$

5.2.3. A more intuitive formulation

As the authors describe in their work, we can consider the Equation (5.1) as a **min-max problem**, that is, an optimization with two steps.

- First, **the minimization of the weighted risk**, $\min_\theta \frac{1}{n}\sum_{i=1}^{n} p_i l_i(\theta, x_i)$, where θ are the parameters to be computed, $\mathbf{x}$ is a set of n samples and p_i is the weight associated to each sample, so the samples with higher contribution to the loss function are the more valuable in the model.
- Second, **the maximization of the robust objective**, $\max_p \sum_{i=1}^{n} p_i l_i$.

As a constraint, they propose the Equation (5.2), where ρ is a parameter to select the confidence level. In the case that p_i is equal to $1/n$ for every sample, the model would correspond to the empirical risk minimization, that is, all the samples have the same weight and indeed, the same contribution.

$$p \in \mathcal{P}_n = \left\{ p \in \mathbb{R}_+^n : \frac{1}{2}\|np - \mathbf{1}\|_2^2 \leq \rho, \langle \mathbf{1}, p \rangle = 1 \right\} \tag{5.2}$$

They give a number of theoretical guarantees and empirical evidences in order to show the optimal performance of the estimator with faster rates of converge and the improvement of out-of-sample test performance in different classification problems.

5.2.4. Application on Deep Learning

Nowadays, deep learning is known as a powerful framework for supervised learning [2]. It allows the implementation of neural networks with as many layers and units as it is desired, providing a more or less sophisticated function to fit a specific dataset. The description of such algorithms is followed by the specification of a cost function, an optimization procedure and a model, what makes the robust objective proposed a direct application in the step of the risk minimization of this kind of tools.

Moreover, neural networks sometimes require long training times when the graph architecture is some how complex. These methods require the use of all the data before updating the predictor. As a consequence, a small improvement at each iteration in the optimization could make huge differences in the performance at the end.

In spite of the high capacity for the adaptation to complex models that deep neural networks have, they involve an excessive computational complexity that makes impossible to apply directly the two-step algorithm from [245] summarized in Section 5.2.1. Evaluating gradients twice would imply go over the entire dataset twice per epoch. We have modified the algorithm in [245] so its computational complexity when training neural networks is negligible compared to uniform sampling.

We would like to express the contribution of the variance to the upper bound of the empirical risk as a way of selecting more frequently the samples with more variance in the mini-batches of the neural networks. This is equivalent to, in the step of computing the gradients, use the worse performed samples more times. With that choice we would like to sacrifice the common classes at better performance on the rare ones.

5.3. Model description

This section describes the methods to select the samples of the mini-batch in a deep learning problem based on the idea described before. We propose four different algorithms.

In the first algorithm, we train the neural network in such a way that, at each iteration, we repeat a percentage of the worst performed samples from the previous mini-batch. This percentage will be a hyper-parameter and has a similar role as the parameter ρ in Equation (5.2), since it lets more samples to have more contribution. We refer to this model as **Variance Reducer per Mini-batch** (VR-M) and it is described in Algorithm 1.

In the second algorithm, we modify the original training set for each epoch so that we repeat a percentage of worst performed samples from the training of the whole previous epoch. This method is detailed in Algorithm 2 and we defined it as **Variance Reducer per Epoch** (VR-E). In this context, we must differ the connotation in the definition of iteration, what we mean as the step from a mini-batch to the next one, with respect to a step between two epochs, which includes many iterations.

Algorithm 1 Variance Reducer per Mini-batch (VR-M)

Require: Datasets $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$.

Ensure: Test accuracy $\boldsymbol{\eta}$.

Initialize parameters $\boldsymbol{\theta}$, number of epochs E and repetition rate ϵ;
for $e = 1 \dots E$ **do**
    Divide dataset $\mathcal{D}_{\text{train}}$ in M mini-batches;
    for $m = 1 \dots M$ **do**
        $\{\mathbf{x}_m, \mathbf{y}_m\} \leftarrow$ Obtain next mini-batch m;
        $\boldsymbol{\ell}_m \leftarrow$ Evaluate cross-entropy in mini-batch m;
        $\boldsymbol{\theta} \leftarrow$ Update parameters with Stochastic Gradient Descent (SGD);
        $\{\mathbf{x}_{m+1}, \mathbf{y}_{m+1}\} \leftarrow$ Substitute $\epsilon \cdot M$ samples with the $\{\mathbf{x}_m, \mathbf{y}_m\}$ of highest $\boldsymbol{\ell}_m$;
    $\eta_e \leftarrow$ Compute test accuracy on $\mathcal{D}_{\text{test}}$;
    Shuffle $\mathcal{D}_{\text{train}}$;

While with the *VR-M* we can repeat a sample almost every iteration, with the *VR-E* we restrict the number of times that a sample is repeated in the overall training because we have much fewer epochs than iterations.

We modify these two algorithms by not including all the samples but only a subset of them. We apply a sampling step with a 50% of random data points belonging to the selection of the top-ranking worst performed ones. This approach helps the method not to insist always on the same samples (which could degrade the quality of the system) and makes the model more robust. That is, if there is a sample that is misleading the method, we could avoid its permanent contribution to the gradients with this solution. In order to make reference to both scenarios it is used **Probabilistic Variance Reducer per Mini-batch** (PVR-M) and **Probabilistic Variance Reducer per Epoch** (PVR-E) respectively for the first and the second model.

Making more clear the differences between this last approach and the basic one, we are exposing an example. Therefore, if in the first algorithm it is repeated at each mini-batch the 40 samples with higher value in the lost function, with the probabilistic approach it would be repeated 20 random samples from these 40 ones.

The Figure 5.1 shows an histogram with the number of times that a sample is used in the optimization. We compare the baseline, that is the original model without repeating any sample, with the two models and their probabilistic approaches. In order to have similar scenarios, we use a repetition of 20% of the samples in the basic versions and 40% with the probabilistic approaches. That is because the latter is resampled half its size so we retain just a quantity of 20% of repeated samples at the end. Indeed, this idea is appreciated better in the VR-E, with almost the same distribution of repeated samples, green and red color bars in the graph. Moreover, we can observe the idea mentioned before, that is, with the probabilistic approach we do not let the model to repeat a sample too many times, as it could happen with the model in yellow with a contribution of almost 3500 times from

Algorithm 2 Variance Reducer per Epoch (VR-E)

Require: Datasets $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$.

Ensure: Test accuracy $\boldsymbol{\eta}$.

Initialize parameters $\boldsymbol{\theta}$, number of epochs E and repetition rate ϵ;
Initialize $\mathcal{D}_{\text{train}}^{(1)} = \mathcal{D}_{\text{train}}$;
for $e = 1 \dots E$ **do**
    Divide dataset $\mathcal{D}_{\text{train}}^{(e)}$ in M mini-batches;
    for $m = 1 \dots M$ **do**
        $\{\mathbf{x}_m, \mathbf{y}_m\} \leftarrow$ Obtain next mini-batch m;
        $\boldsymbol{\ell}_m^{(e)} \leftarrow$ Evaluate cross-entropy in mini-batch m;
        $\boldsymbol{\theta} \leftarrow$ Update parameters with Stochastic Gradient Descent (SGD);
    $\eta_e \leftarrow$ Compute test accuracy on $\mathcal{D}_{\text{test}}$;
    Shuffle $\mathcal{D}_{\text{train}}$;
    $\mathcal{D}_{\text{train}}^{(e+1)} \leftarrow \mathcal{D}_{\text{train}}$;
    $\mathcal{D}_{\text{train}}^{(e+1)} \leftarrow$ Substitute $\epsilon \cdot E$ samples with $\{x_i, y_i\} \in \mathcal{D}_{\text{train}}^{(e)}$ of highest $\boldsymbol{\ell}^{(e)}$;

a set of samples. The baseline defines the number of iterations of the model, 500, that is the number of epochs, since a sample contributes one time per epoch in a original deep learning algorithm.

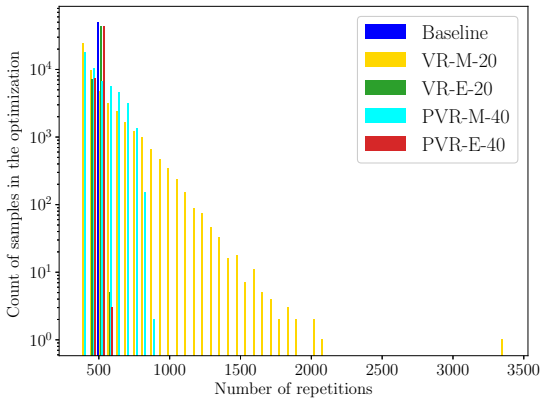


Figure 5.1: Histogram with the number of repetitions of the samples in the CIFAR-10 dataset with the all-CNN architecture. It is used a mini-batch of 128 samples and a dropout of 0.5. It is compared the percentages of 20 and 40 for both the model and the probabilistic approach.

5.4. Experiments

5.4.1. Model

We trained our method in a images classification problem through several scenarios in order to generalize its properties. In consequence, we studied different datasets and networks from the literature.

Datasets

Between all the available datasets, it has been chosen the benchmarks MNIST, SVHN and CIFAR-10 due to their multiple appearances in state-of-the-art works. They allow an easier and faster training of the experiments in comparison with larger bases as the ImageNet [248]. For this reason, the extension of this work in more complex domains will remain as a future task.

The MNIST is composed of 60000 training samples and 10000 test samples of hand-written digits [249]. The images are of size 28x28 pixels in gray scale. It is the simpler dataset used in this work.

The SVHN consists on 32-by-32 RGB images of house numbers from Google Street View [250]. It has 73257 digits for training and 26032 digits for testing.

Finally, the CIFAR-10 collects labeled images of 10 classes [251]. They are 32x32 color pixels and a total of 50000 samples for training and 10000 for test.

Architectures

As it was mentioned in the section 5.2.4, we are proving the behavior of our method in CNNs. For this purpose, we are studying two different architectures of networks from the literature adapted to the datasets mentioned in the previous section.

The first one is based on the VGG implemented by [252]. The motivation of this choice is the validation of the method in a complex enough neural network where the improvements are considerably more cost efficient. The original architecture has been modified according to the size of our data, resulting in a neural network of 11 layers. It has three levels, the first one with two convolutional layers of output 16, the second with other two of output 32 and the third with four layers of output 64. All levels are ended with a max-pooling and finally it is applied three fully connected layers of size 1024, except the last one, with size the number of classes. This scheme is resumed in Table 5.1a.

The second one is the network All-CNN-C from [11]. This particular architecture replaces the max-pooling choice by convolutional layers with increased stride as shown in Table 5.1b. In the training of this scheme, it has been used an adaptive learning rate as in the original work.

Input image
Conv 3x3-16 (with ReLu)
Conv 3x3-16 (with ReLu)
Max-pooling 2x2
Conv 3x3-32 (with ReLu)
Conv 3x3-32 (with ReLu)
Max-pooling 2x2
Conv 3x3-64 (with ReLu)
Conv 3x3-64 (with ReLu)
Conv 3x3-64 (with ReLu)
Conv 3x3-64 (with ReLu)
Max-pooling 2x2
Fully-Connected 1024 (with ReLu)
Dropout 0.5
Fully-Connected 1024 (with ReLu)
Dropout 0.5
Fully-Connected #classes
Soft-max

(a) VGG11b based on the VGG of 11 layers [252].

Input image
Dropout 0.8
Conv 3x3-96 (with ReLu)
Conv 3x3-96 (with ReLu)
Conv 3x3-96 (with ReLu) stride r=2
Dropout 0.5
Conv 3x3-192 (with ReLu)
Conv 3x3-192 (with ReLu)
Conv 3x3-192 (with ReLu) stride r=2
Dropout 0.5
Conv 3x3-192 (with ReLu)
Conv 1x1-192 (with ReLu)
Conv 1x1-#classes (with ReLu)
Global averaging over 6x6 spatial dimensions
Soft-max

(b) All-CNN-C [11].

All the experiments have been trained with *tensorflow*.

5.4.2. Results

The results shown in this section are trained through 200 or 500 epochs with different distributions of the train and test sets, so we can notice one of the advantages of our work in a scenario with less training images. In the cases where we reduce the number of training samples, those ones that are removed are included in the validation set, so it will not be convenient to compare scores with different number of training images. In Tables 5.2, 5.3 and 5.4 we resume the validation accuracy of different configurations and we remark in bold the scores that overcome the baseline and in red the best choice among all.

In the case of the MNIST dataset, we used the VGG with 11 layers as described in Table 5.1a. The mini-batch size was set to 64, the learning rate 0.001 and the initialization of the parameters was 0.1 for the standard deviation of the weights and 0 for the biases. 200 epochs were enough for all the scenarios to converge except for the one with 1000 training samples that we used 500 epochs. Table 5.2 resumes the validation accuracy for different number of training samples, from the original configuration, 60000 training images, until 1000. In addition, we wanted to check the behavior of our method without dropout, what we have called '30000 DP1' in the table, since we used 30000 training samples and set dropout probability to 1, that is the same as removing it from the architecture.

We can confirm from Table 5.2 that our model overcomes the baseline when the number of samples is reduced in any quantity, even when we do not use another regularization mechanism as it is dropout. In addition, we can state through not shown tests that we also obtain the same improvements in accuracy in other scenarios without dropout. Regarding this dataset, the best model is the VR-M, although the advantages are obtained with both approaches.

Table 5.2: Validation accuracy on MNIST with the VGG11 based network.

Models	# training samples							
	60000	50000	40000	30000	20000	10000	1000	30000 DP1
Baseline	99.760%	99.199%	99.046%	98.921%	98.626%	98.319%	93.934%	98.128%
VR-M-5	99.599%	**99.432%**	99.018%	**99.044%**	**98.802%**	98.181%	**95.086%**	97.899%
VR-M-10	99.619%	**99.312%**	**99.099%**	**99.064%**	**98.722%**	98.259%	93.727%	**98.134%**
VR-M-15	99.659%	**99.299%**	**99.207%**	**99.030%**	**98.844%**	**98.416%**	**94.420%**	**98.217%**
VR-M-20	99.659%	**99.406%**	**99.123%**	**99.061%**	**98.940%**	**98.414%**	93.411%	**98.154%**
PVR-M-10	99.579%	**99.346%**	**99.203%**	**98.953%**	**98.809%**	98.254%	**94.638%**	**98.355%**
PVR-M-20	99.619%	**99.332%**	**99.139%**	**99.036%**	**98.722%**	98.245%	**94.258%**	**98.177%**
PVR-M-30	99.700%	**99.272%**	**99.163%**	**99.116%**	**98.829%**	**98.463%**	93.616%	98.114%
PVR-M-40	99.599%	**99.306%**	**99.187%**	**98.998%**	**98.800%**	**98.443%**	**94.272%**	98.060%
VR-E-10	99.599%	**99.232%**	**99.123%**	98.855%	**98.691%**	98.248%	**94.752%**	**98.140%**
VR-E-20	99.679%	**99.359%**	**99.111%**	**98.978%**	**98.637%**	98.142%	**94.291%**	**98.211%**
PVR-E-20	99.639%	**99.319%**	**99.099%**	**98.998%**	**98.729%**	98.250%	92.989%	97.897%
PVR-E-40	99.679%	**99.272%**	**99.091%**	98.884%	**98.717%**	98.172%	93.905%	**98.292%**

Figure 5.2a exposes another interest of our method besides the improvement in the accuracy. That is the faster convergence. In the figure we can differentiate in blue the baseline that goes below the rest of the curves (a set of configurations from our method in the scenario with 60000 training samples) during the first epochs, until the number 55 approximately. After that, the convergence of the baseline follows a better score that the other ones. Then, we could take advantage of this result to apply our method just in the first stage of the training in a particular problem so we can speed up the convergence.

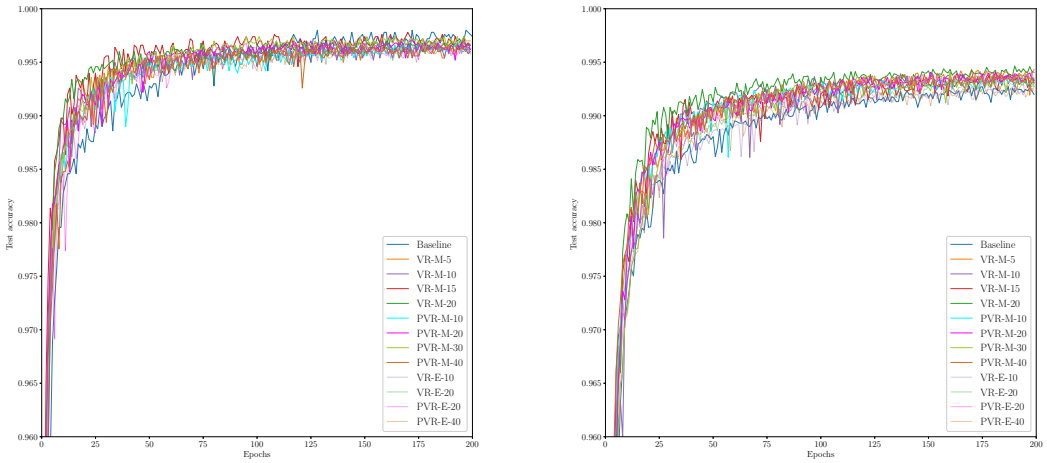
Figure 5.2b shows the accuracy evolution with 50000 training samples, where our method works quite well, maintaining the baseline curve with the worst score during almost the complete training of the algorithm, what was expected from results in Table 5.2.

Figure 5.3 shows the results for the experiments without dropout. In this case it is more visible the differences between the convergence of the baseline and our proposal methods. Precisely, the variance that each curve presents during the training is lower, what allows us to see them quite clear and distanced.

In the same way, we have trained the SVHN dataset with the same network and configuration than the MNIST, but with a batch size of 128 and a fix number of 500 epochs. The results are collected in Table 5.3.

The accuracy improvements for the classification of the images in the SVHN dataset are not so successful as with MNIST but even though we can appreciate some advantages. That is the case of the scenario with the less training samples, 10000, where we overcome until almost 5% the baseline score.

Finally, the training of the CIFAR-10 dataset is studied with the all-CNN from Table 5.1b. We employed a batch size of 128 samples and the adaptive learning rate with the



(a) 60000 training images.

(b) 50000 training images.

Figure 5.2: Validation accuracy per epoch in the MNIST dataset with the VGG11b architecture. It is used a mini-batch of 64 samples and a dropout of 0.5. It is compared the percentages of samples repetition as detailed in Table 5.2.

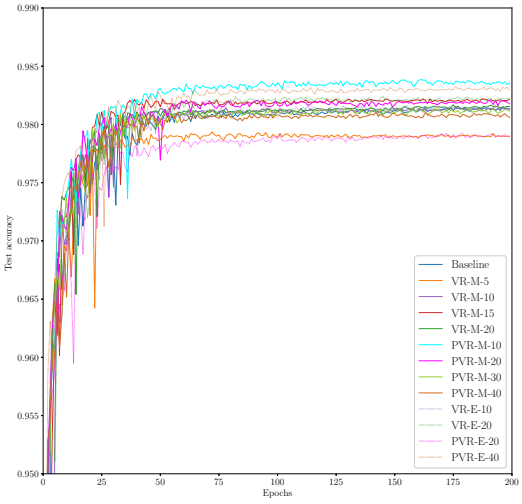


Figure 5.3: Validation accuracy per epoch in the MNIST dataset with the VGG11b architecture. It is used a mini-batch of 64 samples and a dropout of 1. It is compared the percentages of samples repetition as detailed in Table 5.2.

initial value of 0.01. The initialization of the parameters has been set to a standard deviation of 0.05 for the weights and 0 for the biases. In order to increase the baseline accuracy, we have applied a preprocessing step to the images that consists on a global contrast normalization and a ZCA whitening following [253]. The accuracies are exposed in Table

Table 5.3: Validation accuracy on SVHN with the VGG11 based network.

Models	# training samples				
	73257	60000	30000	20000	10000
Baseline	92.064%	91.753%	90.069%	88.787%	80.938%
VR-M-5	91.649%	91.501%	89.130%	86.933%	**84.141%**
VR-M-10	90.602%	91.044%	88.258%	87.196%	80.813%
VR-M-15	91.364%	90.597%	88.254%	85.713%	**81.606%**
VR-M-20	90.144%	90.403%	87.701%	86.561%	**83.300%**
PVR-M-10	**92.372%**	91.557%	89.474%	87.528%	**83.941%**
PVR-M-20	91.918%	**92.034%**	89.542%	88.185%	**85.252%**
PVR-M-30	**92.438%**	91.621%	89.497%	88.526%	**83.444%**
PVR-M-40	**93.003%**	91.577%	89.688%	88.060%	**85.246%**
VR-E-10	**92.153%**	91.590%	89.760%	88.559%	**84.522%**
VR-E-20	**92.330%**	**92.136%**	**90.098%**	87.901%	**81.424%**
PVR-E-20	91.687%	**92.175%**	89.613%	88.501%	**85.311%**
PVR-E-40	**92.403%**	**91.804%**	89.200%	88.559%	**84.168%**

5.4.

With the study of this network, we can discover another possible advantage of our work in the configurations with less training samples. In Table 5.4, when we decrease the number of training samples, our method works better and more cases that overcome the baseline appear. Moreover, the differences in the score between the baseline and the others are higher, so it has more sense the use of our approach with important improvement on the accuracy. That is something that happened in the SVHN dataset with 10000 images, but this time in higher proportion with more than a 6% of increase in the accuracy with 5000 training images. Therefore, our method could be very useful when the number of samples is not high enough.

Regarding the choice of the percentage of repetition in the samples, we do not expose any evidence of trend that it may follow according to the number of training samples or the complexity of the network. Consequently, we should try different alternatives to find the best hyper-parameter. We just advise not to use very high percentages that would remove the sense of the method. Nevertheless, we found that the VR-M works better in datasets as MNIST and CIFAR-10, while in the case of SVHN, with the worse contribution of our method, VR-E appears to works better than the VR-M.

The code to launch the simulations from this section is released on GitHub[13] For its

[13] github.com/AuroraCoboAguilera/RobustSampling

Table 5.4: Validation accuracy on CIFAR-10 with the all-CNN.

	# TRAINING SAMPLES					
MODELS	50000	40000	30000	20000	10000	5000
BASELINE	88.131%	**87.720%**	85.457%	82.545%	76.328%	69.360%
VR-M-5	87.981%	87.380%	**85.763%**	**83.246%**	76.232%	69.262%
VR-M-10	88.041%	87.685%	85.266%	**83.777%**	**77.434%**	**69.549%**
VR-M-15	87.871%	87.565%	85.403%	**82.562%**	**76.899%**	**76.520%**
VR-M-20	88.061%	87.009%	**85.677%**	82.537%	**76.825%**	68.160%
PVR-M-10	**88.331%**	87.309%	**85.991%**	**82.559%**	**76.400%**	**69.675%**
PVR-M-20	88.131%	87.319%	85.153%	**83.188%**	**77.742%**	**70.161%**
PVR-M-30	87.971%	87.354%	84.372%	**83.213%**	**77.220%**	**69.537%**
PVR-M-40	88.021%	87.650%	84.696%	**82.762%**	**77.003%**	**69.639%**
VR-E-10	88.021%	87.019%	85.410%	**82.933%**	76.117%	69.318%
VR-E-20	87.720%	87.405%	**85.577%**	**82.802%**	**76.446%**	69.169%
PVR-E-20	87.971%	86.899%	**85.557%**	**83.108%**	75.663%	**69.668%**
PVR-E-40	87.821%	86.859%	85.123%	82.379%	**76.512%**	**69.974%**

implementation we used the library of *tensorflow*[14].

5.5. Conclusions

In this work we have presented a novel idea for the selection of samples in the training of a deep learning model, based in the variance reduction of the real risk. It consists on the simple idea of repeating the samples with higher variance that are the ones with worse score in the cost function.

We propose several models and study their performance in different architectures and datasets. We discuss the advantages according to the studied problem. Between them, we show the improvement of the accuracy in the classification when the number of training samples is low and the faster rates of convergence. However, we do not expose any evidence for the choice of the value of the percentage hyper-parameter, what has to be tested in the problem to solve. Finally, we highlight the use of our work without dropout, with greater differences in the convergence accuracy and a more statistical relevant increase of the score.

[14] tensorflow.org

6. CONCLUSIONS

The preceding chapters comprise the 4 areas of study during the current doctoral thesis. In this last chapter, we present a summary collecting the principal results and contributions on each area as well as the future lines of research that these studies let open.

6.1. Summary and contributions

The treatment of mental disorders nowadays entails a wide variety of still non-solved tasks such as misdiagnosis or delayed diagnosis. During this doctoral thesis we study and develop different models that can serve as potential tools for the clinician labor. Among our proposals, we outline two main lines of research, Natural Language Processing and probabilistic methods.

In Chapter 2, we start our thesis with a regularization mechanism used in language models and specially effective in Transformer-based architectures, where we call it NoR-BERT, from Noisy Regularized Bidirectional Representations from Transformers [9], [15]. According to the literature, we found out that regularization in NLP is a low explored field limited to the use of general mechanisms such as dropout [57] or early stopping [58]. In this landscape, we propose a novel approach to combine any LM with Variational Auto-Encoders [23]. VAEs belong to deep generative models, with the construction of a regular latent space that permits the reconstruction of the input samples throughout an encoder and decoder networks. Our VAE is based in a prior distribution of a mixture of Gaussians (GMVAE), what gives the model the chance to capture some multimodal information. Combining both, Transformers and GMVAEs we build an architecture capable of imputing missing words from a text corpora in a diverse topic space as well as improve BLEU score in the reconstruction of the data base. Both results depend on the depth of the regularized layer from the Transformer Encoder. The regularization in essence is formed by the GMVAE reconstruction of the Transformer embeddings at some point in the architecture, adding structure noise that helps the model a better generalization. We show improvements in BERT[15], RoBERTa [16] and XLM-R [17] models, verified in different datasets and we also provide explicit examples of sentences reconstructed by Top NoRBERT. In addition, we validate the abilities of our model in data augmentation, improving classification accuracy and F_1 score in various datasets and scenarios thanks to augmented samples generated by NoRBERT. We study some variations in the model, Top, Deep and contextual NoRBERT, the latter based in the use of contextual words to reconstruct the embeddings in the corresponding Transformer layer.

We continue with the Transformers line of research in Chapter 3, proposing PsyBERT. PsyBERT, as the own name refers, is a BERT-based [15] architecture suitably modified to work in Electronic Health Records from psychiatry patients. It is inspired by BEHRT

[19], also devoted to EHRs in general health. We distinguish our model from the training methodology and the embedding layer. In a similar way that with NoRBERT, we find the utility of using a Masked Language Modeling (MLM) policy without no finetuning or specific-task layer at all. On the one hand, we used MLM in NoRBERT to solve the task of imputing missing words, finishing the aim of the model in generating new sentences by inputs with missing information. On the other hand, we firstly propose the use of PsyBERT such as tool to fill the missing diagnoses in the EHR as well as correct misdiagnosed cases. After this task, we also apply PsyBERT in delusional disorder detection. On the contrary, in this scenario we apply a multi-label classification layer, that aims to compute the probability of the different diagnoses in the last visit of the patient to the hospital. From these probabilities, we analyse delusional cases and propose a tool to detect potential candidates of this mental disorder. In both tasks, we make use of several fields obtained from the patient EHR, such as age, sex, diagnoses, treatments of psychiatric history and propose a method capable of combining heterogeneous data to help the diagnosis in mental health. During these works, we point out the problematic in the quality of the data from the EHRs [104], [105] and the great advantage that medical assistance tools like our model can provide. We do not only solve a classification problem with more than 700 different illnesses, but we bring a model to help doctors in the diagnosis of very complex scenarios, with comorbidity, long periods of patient exploration by traditional methodology or low prevalence cases. We present a powerful method treating a problematic with great necessity.

Following the health line of research and psychiatry application, we analyse in Chapter 4 a probabilistic method to search for behavioral pattern in patients also with mental disorders. In this case it is not the method the contribution of the work but the application and results in collaboration with the clinician interpretation. The model is called SPFM (Sparse Poisson Factorization Model) [22] and consist on a non-parametric probabilistic model based on the Indian Buffet Process (IBP) [20], [21]. It is a exploratory method capable of decomposing the input data in sparse matrixes. For that, it imposes the Poisson distribution to the product of two matrixes, Z and B, both obtained respectively by the IBP and a Gamma distribution. Hence Z corresponds to a binary matrix representing active latent features in a patient data and B weights the contribution of the data characteristics to the latent features. The data we use in the three works described during the chapter refers to different questions from e-health questionnaries. Then, the data characteristics refer to the answer or punctuation on each question and the latent features from different behavioral patterns in a patient regarding the selection of features active in their questionnaires. For example, patient X can present feature 1 and 2 and patient Y may presence feature 1 and 3, giving as a result two different profiles of behavioral. With these procedure we study three scenarios. In the first problematic, we relate the profiles with the diagnoses, finding common patterns among the patients and connections between diseases. We also analyse the grade of critical state and contrast the clinician judgment via the Clinical Global Impression (CGI). In the second scenario, we pursue a similar study and find out connections between disturbed sleeping patterns and clinical markers of wish to die.

We focus this analysis in patients with suicidal thoughts due to the problematic that those individuals suppose as a major public health issue [175]. In this case we vary the questionnarie and the data sample, obtaining different profiles also with important information to interpret by the psychiatrist. The main contribution of this work is the proportion of a mechanism capable of helping with detection and prevention of suicide. Finally, the third work comprehend a behavioral pattern study in mental health patient before and during covid-19 lockdown. We did not want to lose the chance to contribute during coronavirus disease outbreak and presented a study about the changes in psychiatric patients during the alarm state. We analyse again the profiles with the previous e-health questionnaire and discover that the self-reported suicide risk decreased during the lockdown. These results contrast with others studies [237] and suppose signs for an increase in suicidal ideation once the crisis ceases.

Finally, Chapter 5 propose a regularization mechanism based in a theoretical idea from [245] to obtain a variance reduction in the real risk. We interpret the robust regularized risk that those authors propose in a two-step mechanism formed by the minimization of the weighted risk and the maximization of a robust objective and suggest an idea to apply this methodology in a way to select the samples from the mini-batch in a deep learning set up. We study different variations of repeating the worst performed samples from the previous mini-bath during the training procedure and show proves of improvements in the accuracy and faster convergence rates of a image classification problem with different architectures and datasets.

6.2. Future lines of research

We would like to emphasize two difficulties found during the doctoral thesis and some lines of research according to them.

On the one hand, we focus in NLP models. Nowadays, there is no robust metric to evaluate the performance of a language generative model as the one we propose. Usually, we validate the results in other tasks, such as the data augmentation for other task solution or the GLUE [86] proposition in the literature. We consider that there is an open line of research in terms of evaluation metrics capable of collecting semantic and syntactic properties to test text generation models.

On the other hand, we discover with the help of clinicians, an difficult problem to solve with today's EHRs due to the great quantities of missing information, the non encoded sources, and the incorrect values [104], [106], [108]. Even though we propose an useful tool to combine with expertise opinion in the decision making, we find it a very general issue translated to other diseases to be explored and not only psychiatry. In medicine, some illnesses diagnosis may require long periods of observation maybe resulting in a late diagnosis. With ML we can help the doctors detect determined behavioral patterns or indicators that detect a suspicious disease and avoid a critical end. We keep an open line to

be research with the improvement of the present models or the development of new ones with the advance of technologies.

Finally, we would like to mention the combination of results from chapters 2 and 3, with a regularized version of PsyBERT. We consider that this could be an interesting work to do in the future and a continuation of the presented doctoral thesis. With this idea we could take advantage of the power from probabilistic models, the scope from Transformers and the information source from heterogeneous data that could result in a practical method.

BIBLIOGRAPHY

[1] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[4] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[5] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*, PMLR, 2017, pp. 214–223.

[6] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[7] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[8] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, PMLR, 2016, pp. 1050–1059.

[9] A. Vaswani *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[10] S. H. Jain, B. W. Powers, J. B. Hawkins, and J. S. Brownstein, "The digital phenotype," *Nature biotechnology*, vol. 33, no. 5, pp. 462–463, 2015.

[11] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[12] H. Namkoong, A. Sinha, S. Yadlowsky, and J. C. Duchi, "Adaptive sampling probabilities for non-smooth optimization," in *International Conference on Machine Learning*, 2017, pp. 2574–2583.

[13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[14] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1412–1421.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[16] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[17] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[18] N. Dilokthanakul *et al.*, "Deep unsupervised clustering with gaussian mixture variational autoencoders," *arXiv preprint arXiv:1611.02648*, 2016.

[19] Y. Li *et al.*, "Behrt: Transformer for electronic health records," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[20] T. L. Griffiths and Z. Ghahramani, "The indian buffet process: An introduction and review.," *Journal of Machine Learning Research*, vol. 12, no. 4, 2011.

[21] M. Chen, C. Gao, and H. Zhao, "Posterior contraction rates of the phylogenetic indian buffet processes," *Bayesian analysis*, vol. 11, no. 2, p. 477, 2016.

[22] M. F. Pradier, V. Stojkoski, Z. Utkovski, L. Kocorev, and F. Perez-Cruz, "Sparse three-parameter restricted indian buffet process for understanding international trade," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 2476–2480.

[23] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *stat*, vol. 1050, p. 1, 2014.

[24] M. Caccia, L. Caccia, W. Fedus, H. Larochelle, J. Pineau, and L. Charlin, "Language gans falling short," in *Proceedings of the Eighth International Conference on Learning Representations, ICLR 2020*, 2020.

[25] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International Conference on Machine Learning*, PMLR, 2019, pp. 7354–7363.

[26] T. Yang, P. Ren, X. Xie, and L. Zhang, "Gan prior embedded network for blind face restoration in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 672–681.

[27] N.-T. Tran, V.-H. Tran, N.-B. Nguyen, T.-K. Nguyen, and N.-M. Cheung, "On data augmentation for gan training," *IEEE Transactions on Image Processing*, vol. 30, pp. 1882–1897, 2021.

[28] S. Lu, Y. Zhu, W. Zhang, J. Wang, and Y. Yu, "Neural text generation: Past, present and beyond," *arXiv preprint arXiv:1803.07133*, 2018.

[29] L. Yu, W. Zhang, J. Wang, and Y. Yu, "Seqgan: Sequence generative adversarial nets with policy gradient," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 2852–2858.

[30] W. Fedus, I. Goodfellow, and A. M. Dai, "Maskgan: Better text generation via filling in the _," in *International Conference on Learning Representations*, 2018.

[31] J. Guo, S. Lu, H. Cai, W. Zhang, Y. Yu, and J. Wang, "Long text generation via adversarial training with leaked information," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[32] C. de Masson d'Autume, S. Mohamed, M. Rosca, and J. Rae, "Training language gans from scratch," in *Advances in Neural Information Processing Systems*, 2019, pp. 4302–4313.

[33] E. Jang, S. Gu, and B. Poole, "Categorical reparametrization with gumble-softmax," in *International Conference on Learning Representations (ICLR 2017)*, OpenReview. net, 2017.

[34] Y. Zhang *et al.*, "Adversarial feature matching for text generation," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, JMLR. org, 2017, pp. 4006–4015.

[35] W. Nie, N. Narodytska, and A. Patel, "Relgan: Relational generative adversarial networks for text generation," in *International conference on learning representations*, 2019.

[36] J. Zhao, Y. Kim, K. Zhang, A. M. Rush, and Y. LeCun, "Adversarially regularized autoencoders," in *35th International Conference on Machine Learning, ICML 2018*, International Machine Learning Society (IMLS), 2018, pp. 9405–9420.

[37] S. Subramanian, S. R. Mudumba, A. Sordoni, A. Trischler, A. C. Courville, and C. Pal, "Towards text generation with adversarially learned neural outlines," in *Advances in Neural Information Processing Systems*, 2018, pp. 7551–7563.

[38] D. Donahue and A. Rumshisky, "Adversarial text generation without reinforcement learning," *arXiv preprint arXiv:1810.06640*, 2018.

[39] W. Yu *et al.*, "Learning deep network representations with adversarially regularized autoencoders," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2663–2671.

[40] M. A. Haidar, M. Rezagholizadeh, A. Do Omri, and A. Rashid, "Latent code and text-based generative adversarial networks for soft-text generation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 2248–2258.

[41] M. A. Haidar and M. Rezagholizadeh, "Textkd-gan: Text generation using knowledge distillation and generative adversarial networks," in *Canadian Conference on Artificial Intelligence*, Springer, 2019, pp. 107–118.

[42] A. Rashid, A. Do-Omri, M. A. Haidar, Q. Liu, and M. Rezagholizadeh, "Bilingual-gan: A step towards parallel text generation," in *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, 2019, pp. 55–64.

[43] P. Grnarova, Y. Kilcher, K. Y. Levy, A. Lucchi, and T. Hofmann, "Generative minimization networks: Training gans without competition," *arXiv preprint arXiv:2103.12685*, 2021.

[44] Z. He, M. Kan, and S. Shan, "Eigengan: Layer-wise eigen-learning for gans," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 408–14 417.

[45] Y. Jiang, S. Chang, and Z. Wang, "Transgan: Two pure transformers can make one strong gan, and that can scale up," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[46] A. Pagnoni, K. Liu, and S. Li, "Conditional variational autoencoder for neural machine translation," *arXiv preprint arXiv:1812.04405*, 2018.

[47] D. Shen, Y. Zhang, R. Henao, Q. Su, and L. Carin, "Deconvolutional latent-variable model for text sequence matching," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[48] A. Gupta, A. Agarwal, P. Singh, and P. Rai, "A deep generative framework for paraphrase generation," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[49] Z. Yang, Z. Hu, R. Salakhutdinov, and T. Berg-Kirkpatrick, "Improved variational autoencoders for text modeling using dilated convolutions," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, JMLR. org, 2017, pp. 3881–3890.

[50] W. Shi, H. Zhou, N. Miao, S. Zhao, and L. Li, "Fixing gaussian mixture vaes for interpretable text generation," *arXiv preprint arXiv:1906.06719*, 2019.

[51] C. Li *et al.*, "Optimus: Organizing sentences via pre-trained modeling of a latent space," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4678–4699.

[52] V. Prokhorov, Y. Li, E. Shareghi, and N. Collier, "Learning sparse sentence encoding without supervision: An exploration of sparsity in variational autoencoders," in *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, 2021, pp. 34–46.

[53] S. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," in *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 2016, pp. 10–21.

[54] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Toward controlled generation of text," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, JMLR. org, 2017, pp. 1587–1596.

[55] A. Sriram, H. Jun, S. Satheesh, and A. Coates, "Cold fusion: Training seq2seq models together with language models," *Proc. Interspeech 2018*, pp. 387–391, 2018.

[56] C. Gulcehre *et al.*, "On using monolingual corpora in neural machine translation," *arXiv preprint arXiv:1503.03535*, 2015.

[57] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[58] G. Raskutti, M. J. Wainwright, and B. Yu, "Early stopping and non-parametric regression: An optimal data-dependent stopping rule," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 335–366, 2014.

[59] S. Y. Feng *et al.*, "A survey of data augmentation approaches for nlp," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 968–988.

[60] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in *Advances in neural information processing systems*, 1992, pp. 950–957.

[61] L. Zehui, P. Liu, L. Huang, J. Chen, X. Qiu, and X. Huang, "Dropattention: A regularization method for fully-connected self-attention networks," *arXiv preprint arXiv:1907.11065*, 2019.

[62] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, 2002, pp. 311–318.

[63] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou, "Variational deep embedding: An unsupervised and generative approach to clustering," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 1965–1972.

[64] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in neural information processing systems*, 2015, pp. 3483–3491.

[65] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[66] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2978–2988.

[67] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, "Ctrl: A conditional transformer language model for controllable generation," *arXiv preprint arXiv:1909.05858*, 2019.

[68] X. Ma *et al.*, "A tensorized transformer for language modeling," in *Advances in Neural Information Processing Systems*, 2019, pp. 2229–2239.

[69] J. Gu, C. Wang, and J. Zhao, "Levenshtein transformer," in *Advances in Neural Information Processing Systems*, 2019, pp. 11 179–11 189.

[70] S. Yang *et al.*, "On the localness modeling for the self-attention based end-to-end speech synthesis," *Neural Networks*, vol. 125, pp. 121–130, 2020.

[71] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *unpublished work*, 2018. [Online]. Available: https://s3-us-west-2.%20amazonaws.%20com/openai-assets/researchcovers/languageunsupervised/language%20understanding%20paper..

[72] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mass: Masked sequence to sequence pre-training for language generation," in *International Conference on Machine Learning*, 2019, pp. 5926–5936.

[73] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *unpublished work*, 2019. [Online]. Available: https://s3-us-west-2.amazonaws.%20com/openai-assets/research-covers/language-unsupervised/%20language_understanding_paper.pdf..

[74] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in neural information processing systems*, 2019, pp. 5754–5764.

[75] G. Lample and A. Conneau, "Cross-lingual language model pretraining," *arXiv preprint arXiv:1901.07291*, 2019.

[76] L. Dong *et al.*, "Unified language model pre-training for natural language understanding and generation," in *Advances in Neural Information Processing Systems*, 2019, pp. 13 042–13 054.

[77] Y. Sun *et al.*, "Ernie 2.0: A continual pre-training framework for language understanding," in *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[78] H. Xiao, "Hungarian layer: A novel interpretable neural layer for paraphrase identification," *Neural Networks*, vol. 131, pp. 172–184, 2020.

[79] M. Yang, L. Chen, Z. Lyu, J. Liu, Y. Shen, and Q. Wu, "Hierarchical fusion of common sense knowledge and classifier decisions for answer selection in community question answering," *Neural Networks*, vol. 132, pp. 53–65, 2020.

[80] M. Zaheer *et al.*, "Big bird: Transformers for longer sequences," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 283–17 297, 2020.

[81] J. Su, Y. Lu, S. Pan, B. Wen, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *arXiv preprint arXiv:2104.09864*, 2021.

[82] X. Ouyang *et al.*, "Ernie-m: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 27–38.

[83] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *arXiv preprint arXiv:2101.03961*, 2021.

[84] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[85] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *International Conference on Learning Representations*, 2020.

[86] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," in *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[87] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[88] Y. Zhu *et al.*, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 19–27.

[89] H. [dataset] Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[90] D. [dataset] Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30k: Multilingual english-german image descriptions," in *Proceedings of the 5th Workshop on Vision and Language*, 2016, pp. 70–74.

[91] E. Loper and S. Bird, "Nltk: The natural language toolkit," in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, 2002, pp. 63–70.

[92] S. [dataset] Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 632–642.

[93] R. [dataset] Socher *et al.*, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631–1642.

[94] E. M. [dataset] Voorhees and D. M. Tice, "Building a question answering test collection," in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 2000, pp. 200–207.

[95] O. Kovaleva *et al.*, "Revealing the dark secrets of bert," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, American Medical Informatics Association, vol. 1, 2019, pp. 2465–2475.

[96] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in neural information processing systems*, 2017, pp. 6626–6637.

[97] C.-Y. Lin and F. Och, "Looking for a few good metrics: Rouge and its evaluation," in *Ntcir Workshop*, 2004.

[98] B. Wang, A. Wang, F. Chen, Y. Wang, and C.-C. J. Kuo, "Evaluating word embedding models: Methods and experimental results," *APSIPA Transactions on Signal and Information Processing*, vol. 8, 2019.

[99] J. P. Pestian *et al.*, "A machine learning approach to identifying the thought markers of suicidal subjects: A prospective multicenter trial," *Suicide and Life-Threatening Behavior*, vol. 47, no. 1, pp. 112–121, 2017.

[100] S. Velupillai *et al.*, "Risk assessment tools and data-driven approaches for predicting and preventing suicidal behavior," *Frontiers in psychiatry*, vol. 10, p. 36, 2019.

[101] N. Rezaii, E. Walker, and P. Wolff, "A machine learning approach to predicting psychosis using semantic density and latent content analysis," *NPJ schizophrenia*, vol. 5, no. 1, pp. 1–12, 2019.

[102] G. Fond *et al.*, "Machine learning for predicting psychotic relapse at 2 years in schizophrenia in the national face-sz cohort," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 92, pp. 8–18, 2019.

[103] C. Crema, G. Attardi, D. Sartiano, and A. Redolfi, "Natural language processing in clinical neuroscience and psychiatry: A review," *Frontiers in Psychiatry*, vol. 13, 2022.

[104] K. S. Chan, J. B. Fowles, and J. P. Weiner, "Electronic health records and the reliability and validity of quality measures: A review of the literature," *Medical Care Research and Review*, vol. 67, no. 5, pp. 503–527, 2010.

[105] S. L. Feder, "Data quality in electronic health records research: Quality domains and assessment methods," *Western journal of nursing research*, vol. 40, no. 5, pp. 753–766, 2018.

[106] J. M. Madden, M. D. Lakoma, D. Rusinak, C. Y. Lu, and S. B. Soumerai, "Missing clinical and behavioral health data in a large electronic health record (ehr) system," *Journal of the American Medical Informatics Association*, vol. 23, no. 6, pp. 1143–1149, 2016.

[107] I. Petersen *et al.*, "Health indicator recording in uk primary care electronic health records: Key implications for handling missing data," *Clinical epidemiology*, vol. 11, p. 157, 2019.

[108] B. J. Wells, K. M. Chagin, A. S. Nowacki, and M. W. Kattan, "Strategies for handling missing data in electronic health record derived data," *Egems*, vol. 1, no. 3, 2013.

[109] G. K. Savova *et al.*, "Mayo clinical text analysis and knowledge extraction system (ctakes): Architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[110] G. Gorrell, X. Song, and A. Roberts, "Bio-yodie: A named entity linking system for biomedical text," *arXiv preprint arXiv:1811.04860*, 2018.

[111] Z. Kraljevic *et al.*, "Multi-domain clinical natural language processing with medcat: The medical concept annotation toolkit," *Artificial Intelligence in Medicine*, vol. 117, p. 102 083, 2021.

[112] S. Bacchi, L. Oakden-Rayner, T. Zerner, T. Kleinig, S. Patel, and J. Jannes, "Deep learning natural language processing successfully predicts the cerebrovascular cause of transient ischemic attack-like presentations," *Stroke*, vol. 50, no. 3, pp. 758–760, 2019.

[113] B. D. Wissel *et al.*, "Investigation of bias in an epilepsy machine learning algorithm trained on physician notes," *Epilepsia*, vol. 60, no. 9, e93–e98, 2019.

[114] Z. Xia *et al.*, "Modeling disease severity in multiple sclerosis using electronic health records," *PloS one*, vol. 8, no. 11, e78927, 2013.

[115] K. Takano, M. Ueno, J. Moriya, M. Mori, Y. Nishiguchi, and F. Raes, "Unraveling the linguistic nature of specific autobiographical memories using a computerized classification algorithm," *Behavior Research Methods*, vol. 49, no. 3, pp. 835–852, 2017.

[116] D. G. Clark *et al.*, "Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 2, pp. 113–122, 2016.

[117] H. Wang, Y. Li, M. Hutch, A. Naidech, Y. Luo, *et al.*, "Using tweets to understand how covid-19–related health beliefs are affected in the age of social media: Twitter data analysis study," *Journal of medical Internet research*, vol. 23, no. 2, e26302, 2021.

[118] A. Le Glaz *et al.*, "Machine learning and natural language processing in mental health: Systematic review," *Journal of Medical Internet Research*, vol. 23, no. 5, e15708, 2021.

[119] N. Rezaii, P. Wolff, and B. H. Price, "Natural language processing in psychiatry: The promises and perils of a transformative approach," *The British Journal of Psychiatry*, vol. 220, no. 5, pp. 251–253, 2022.

[120] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-bert: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–13, 2021.

[121] J. Shang, T. Ma, C. Xiao, and J. Sun, "Pre-training of graph augmented transformers for medication recommendation," in *International Joint Conference on Artificial Intelligence*, International Joint Conferences on Artificial IntelligenceThomas. schiex …, 2019.

[122] R. Kodialam, R. Boiarsky, J. Lim, A. Sai, N. Dixit, and D. Sontag, "Deep contextual clinical prediction with reverse distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 249–258.

[123] Y. Li *et al.*, "Hi-behrt: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records," *arXiv preprint arXiv:2106.11360*, 2021.

[124] S. Rao *et al.*, "Targeted-behrt: Deep learning for observational causal inference on longitudinal electronic health records," *arXiv preprint arXiv:2202.03487*, 2022.

[125] W. H. Organization *et al.*, *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*. World Health Organization, 1992.

[126] T. C. Manschreck and N. L. Khan, "Recent advances in the treatment of delusional disorder," *The Canadian Journal of Psychiatry*, vol. 51, no. 2, pp. 114–119, 2006.

[127] F. Edition *et al.*, "Diagnostic and statistical manual of mental disorders," *Am Psychiatric Assoc*, vol. 21, no. 21, pp. 591–643, 2013.

[128] M. Skelton, W. A. Khokhar, and S. P. Thacker, "Treatments for delusional disorder," *Cochrane Database of Systematic Reviews*, no. 5, 2015.

[129] F. Cegla-Schvartzman *et al.*, "Diagnostic stability in bipolar disorder: A follow-up study in 130,000 patient-years," *The Journal of Clinical Psychiatry*, vol. 82, no. 6, p. 36 681, 2021.

[130] N. Palomar-Ciria, F. Cegla-Schvartzman, J.-D. Lopez-Morinigo, H. J. Bello, S. Ovejero, and E. Baca-Garcia, "Diagnostic stability of schizophrenia: A systematic review," *Psychiatry research*, vol. 279, pp. 306–314, 2019.

[131] B. Oldenburg, C. B. Taylor, A. O'Neil, F. Cocker, and L. D. Cameron, "Using new technologies to improve the prevention and management of chronic conditions in populations," *Annu Rev Public Health*, vol. 36, no. 1, pp. 483–505, 2015.

[132] T. Economist, *A digital revolution in health care is speeding up*, 2017.

[133] M. L. Barrigón *et al.*, "Comparative study of pencil-and-paper and electronic formats of ghq-12, who-5 and phq-9 questionnaires," *Revista de Psiquiatría y Salud Mental (English Edition)*, vol. 10, no. 3, pp. 160–167, 2017.

[134] P. Yellowlees, M. M. Burke, S. L. Marks, D. M. Hilty, and J. H. Shore, "Emergency telepsychiatry," *Journal of telemedicine and telecare*, vol. 14, no. 6, pp. 277–281, 2008.

[135] S. Bucci, M. Schwannauer, and N. Berry, "The digital revolution and its impact on mental health care," *Psychology and Psychotherapy: Theory, Research and Practice*, vol. 92, no. 2, pp. 277–297, 2019.

[136] S. Berrouiguet, M. L. Barrigón, J. L. Castroman, P. Courtet, A. Artés-Rodríguez, and E. Baca-García, "Combining mobile-health (mhealth) and artificial intelligence (ai) methods to avoid suicide attempts: The smartcrises study protocol," *BMC psychiatry*, vol. 19, no. 1, pp. 1–9, 2019.

[137] H. E. Payne, C. Lister, J. H. West, and J. M. Bernhardt, "Behavioral functionality of mobile apps in health interventions: A systematic review of the literature," *JMIR mHealth and uHealth*, vol. 3, no. 1, e3335, 2015.

[138] J. Thakkar *et al.*, "Mobile telephone text messaging for medication adherence in chronic disease: A meta-analysis," *JAMA internal medicine*, vol. 176, no. 3, pp. 340–349, 2016.

[139] S. M. Badawy, L. Barrera, M. G. Sinno, S. Kaviany, L. C. O'dwyer, and L. M. Kuhns, "Text messaging and mobile phone apps as interventions to improve adherence in adolescents with chronic health conditions: A systematic review," *JMIR mHealth and uHealth*, vol. 5, no. 5, e7798, 2017.

[140] M. Al-Durra, M.-B. Torio, J. A. Cafazzo, *et al.*, "The use of behavior change theory in internet-based asthma self-management interventions: A systematic review," *Journal of medical Internet research*, vol. 17, no. 4, e4110, 2015.

[141] L. R. Saslow *et al.*, "An online intervention comparing a very low-carbohydrate ketogenic diet and lifestyle recommendations versus a plate method diet in overweight individuals with type 2 diabetes: A randomized controlled trial," *Journal of medical Internet research*, vol. 19, no. 2, e5806, 2017.

[142] J. Naparstek, R. R. Wing, X. Xu, and T. M. Leahey, "Internet-delivered obesity treatment improves symptoms of and risk for depression," *Obesity*, vol. 25, no. 4, pp. 671–675, 2017.

[143] K. Haas, A. Martin, and K. Park, "Text message intervention (teach) improves quality of life and patient activation in celiac disease: A randomized clinical trial," *The Journal of pediatrics*, vol. 185, pp. 62–67, 2017.

[144] Z. Obermeyer and E. J. Emanuel, "Predicting the future—big data, machine learning, and clinical medicine," *The New England journal of medicine*, vol. 375, no. 13, p. 1216, 2016.

[145] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, 2015.

[146] S. Dreiseitl, L. Ohno-Machado, H. Kittler, S. Vinterbo, H. Billhardt, and M. Binder, "A comparison of machine learning methods for the diagnosis of pigmented skin lesions," *Journal of biomedical informatics*, vol. 34, no. 1, pp. 28–36, 2001.

[147] G. Paslakis *et al.*, "Assessment of use and preferences regarding internet-based health care delivery: Cross-sectional questionnaire study," *Journal of medical Internet research*, vol. 21, no. 5, e12416, 2019.

[148] P. Musiat, P. Goldstone, and N. Tarrier, "Understanding the acceptability of e-mental health-attitudes and expectations towards computerised self-help treatments for mental health problems," *BMC psychiatry*, vol. 14, no. 1, pp. 1–8, 2014.

[149] N. Schulze *et al.*, "Differences in attitudes toward online interventions in psychiatry and psychotherapy between health care professionals and nonprofessionals: A survey," *Telemedicine and e-Health*, vol. 25, no. 10, pp. 926–932, 2019.

[150] S. De Rosis and S. Barsanti, "Patient satisfaction, e-health and the evolution of the patient–general practitioner relationship: Evidence from an italian survey," *Health Policy*, vol. 120, no. 11, pp. 1279–1292, 2016.

[151] T. R. Insel, "Digital phenotyping: Technology for a new science of behavior," *Jama*, vol. 318, no. 13, pp. 1215–1216, 2017.

[152] I. Barnett, J. Torous, P. Staples, L. Sandoval, M. Keshavan, and J.-P. Onnela, "Relapse prediction in schizophrenia through digital phenotyping: A pilot study," *Neuropsychopharmacology*, vol. 43, no. 8, pp. 1660–1666, 2018.

[153] A. M. Bernardos, M. Pires, D. Ollé, and J. R. Casar, "Digital phenotyping as a tool for personalized mental healthcare," in *proceedings of the 13th EAI International Conference on pervasive computing technologies for healthcare*, 2019, pp. 403–408.

[154] V. Konok, D. Gigler, B. M. Bereczky, and Á. Miklósi, "Humans' attachment to their mobile phones and its relationship with interpersonal attachment style," *Computers in Human Behavior*, vol. 61, pp. 537–547, 2016.

[155] C. L. Davidson, M. D. Anestis, and P. M. Gutierrez, "Ecological momentary assessment is a neglected methodology in suicidology," *Archives of suicide research*, vol. 21, no. 1, pp. 1–11, 2017.

[156] E. M. Kleiman and M. K. Nock, "Real-time assessment of suicidal thoughts and behaviors," *Current Opinion in Psychology*, vol. 22, pp. 33–37, 2018.

[157] S. Melbye, L. V. Kessing, J. E. Bardram, M. Faurholt-Jepsen, *et al.*, "Smartphone-based self-monitoring, treatment, and automatically generated data in children, adolescents, and young adults with psychiatric disorders: Systematic review," *JMIR mental health*, vol. 7, no. 10, e17453, 2020.

[158] M. L. Barrigón *et al.*, "User profiles of an electronic mental health tool for ecological momentary assessment: Memind," *International journal of methods in psychiatric research*, vol. 26, no. 1, e1554, 2017.

[159] C. W. Topp, S. D. Østergaard, S. Søndergaard, and P. Bech, "The who-5 well-being index: A systematic review of the literature," *Psychotherapy and psychosomatics*, vol. 84, no. 3, pp. 167–176, 2015.

[160] M. del Pilar Sánchez-López and V. Dresch, "The 12-item general health questionnaire (ghq-12): Reliability, external validity and factor structure in the spanish population," *Psicothema*, vol. 20, no. 4, pp. 839–843, 2008.

[161] J. Busner and S. D. Targum, "The clinical global impressions scale: Applying a research tool in clinical practice," *Psychiatry (Edgmont)*, vol. 4, no. 7, p. 28, 2007.

[162] D. F. Alwin, "Feeling thermometers versus 7-point scales: Which are better?" *Sociological Methods & Research*, vol. 25, no. 3, pp. 318–340, 1997.

[163] N. D. Kieruj and G. Moors, "Variations in response style behavior by response scale format in attitude research," *International journal of public opinion research*, vol. 22, no. 3, pp. 320–342, 2010.

[164] K. Mehrotra, C. K. Mohan, and S. Ranka, *Elements of artificial neural networks*. MIT press, 1997.

[165] A. Gómez-Carrillo *et al.*, "How far is clinical assessment from the bullseye? using memind to compare clinical assessment with self-assessment in patients with depression and anxiety diagnosis," *The European Journal of Psychiatry*, vol. 31, no. 4, pp. 158–164, 2017.

[166] S. Berrouiguet, E. Baca-García, S. Brandt, M. Walter, P. Courtet, *et al.*, "Fundamentals for future mobile-health (mhealth): A systematic review of mobile phone and web-based text messaging in mental health," *Journal of medical Internet research*, vol. 18, no. 6, e5066, 2016.

[167] D. Bakker and N. Rickard, "Engagement in mobile phone app for self-monitoring of emotional wellbeing predicts changes in mental health: Moodprism," *Journal of affective disorders*, vol. 227, pp. 432–442, 2018.

[168] L. L. DuBenske *et al.*, "Chess improves cancer caregivers' burden and mood: Results of an ehealth rct.," *Health psychology*, vol. 33, no. 10, p. 1261, 2014.

[169] D. Mangin, J. Parascandalo, O. Khudoyarova, G. Agarwal, V. Bismah, and S. Orr, "Multimorbidity, ehealth and implications for equity: A cross-sectional survey of patient perspectives on ehealth," *BMJ open*, vol. 9, no. 2, e023731, 2019.

[170] A. Wernhart, S. Gahbauer, and D. Haluza, "Ehealth and telemedicine: Practices and beliefs among healthcare professionals and medical students at a medical university," *PloS one*, vol. 14, no. 2, e0213067, 2019.

[171] K. Raychaudhuri and P. Ray, "Privacy challenges in the use of ehealth systems for public health management," in *Emerging Communication Technologies for E-Health and Medicine*, IGI Global, 2012, pp. 155–166.

[172] P. Ware *et al.*, "Using ehealth technologies: Interests, preferences, and concerns of older adults," *Interactive journal of medical research*, vol. 6, no. 1, e4447, 2017.

[173] S. Bucci *et al.*, "Actissist: Proof-of-concept trial of a theory-driven digital intervention for psychosis," *Schizophrenia bulletin*, vol. 44, no. 5, pp. 1070–1080, 2018.

[174] R. E. Anderson, S. H. Spence, C. L. Donovan, S. March, S. Prosser, and J. Kenardy, "Working alliance in online cognitive behavior therapy for anxiety disorders in youth: Comparison with clinic delivery and its role in predicting outcome," *Journal of medical Internet research*, vol. 14, no. 3, e1848, 2012.

[175] *World health organization. (2020). suicide*, https://www.who.int/news-room/fact-sheets/detail/suicide, Retrieved 25 April 2020.

[176] H. S. Wortzel, S. Nazem, N. H. Bahraini, and B. B. Matarazzo, "Why suicide risk assessment still matters," *Journal of Psychiatric Practice®*, vol. 23, no. 6, pp. 436–440, 2017.

[177] P. Saini, D. While, K. Chantler, K. Windfuhr, and N. Kapur, "Assessment and management of suicide risk in primary care.," *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, vol. 35, no. 6, p. 415, 2014.

[178] S. M. Barnes *et al.*, "Moving beyond self-report: Implicit associations about death/life prospectively predict suicidal behavior among veterans," *Suicide and Life-Threatening Behavior*, vol. 47, no. 1, pp. 67–77, 2017.

[179] E. Baca-Garcia *et al.*, "Estimating risk for suicide attempt: Are we asking the right questions?: Passive suicidal ideation as a marker for suicidal behavior," *Journal of affective disorders*, vol. 134, no. 1-3, pp. 327–332, 2011.

[180] J. Suokas, K. Suominen, E. Isometsä, A. Ostamo, and J. Lönnqvist, *Long-term risk factors for suicide mortality after attempted suicide-findings of a 14-year follow-up study*, 2001.

[181] C. Palacio *et al.*, "Identification of suicide risk factors in medellin, colombia: A case-control study of psychological autopsy in a developing country," *Archives of Suicide Research*, vol. 11, no. 3, pp. 297–308, 2007.

[182] R. T. Liu, A. H. Bettis, and T. A. Burke, "Characterizing the phenomenology of passive suicidal ideation: A systematic review and meta-analysis of its prevalence, psychiatric comorbidity, correlates, and comparisons with active suicidal ideation," *Psychological medicine*, vol. 50, no. 3, pp. 367–383, 2020.

[183] D. Ben-Zeev, E. A. Scherer, R. M. Brian, L. A. Mistler, A. T. Campbell, and R. Wang, "Use of multimodal technology to identify digital correlates of violence among inpatients with serious mental illness: A pilot study," *Psychiatric services*, vol. 68, no. 10, pp. 1088–1092, 2017.

[184] E. M. Kleiman, B. J. Turner, S. Fedor, E. E. Beale, J. C. Huffman, and M. K. Nock, "Examination of real-time fluctuations in suicidal ideation and its risk factors: Results from two ecological momentary assessment studies.," *Journal of abnormal psychology*, vol. 126, no. 6, p. 726, 2017.

[185] A. Hadzic *et al.*, "The association of trait impulsivity and suicidal ideation and its fluctuation in the context of the interpersonal theory of suicide," *Comprehensive psychiatry*, vol. 98, p. 152 158, 2020.

[186] V. Vine, S. E. Victor, H. Mohr, A. L. Byrd, and S. D. Stepp, "Adolescent suicide risk and experiences of dissociation in daily life," *Psychiatry research*, vol. 287, p. 112 870, 2020.

[187] N. Hallensleben *et al.*, "Predicting suicidal ideation by interpersonal variables, hopelessness and depression in real-time. an ecological momentary assessment study in psychiatric inpatients with depression," *European Psychiatry*, vol. 56, no. 1, pp. 43–50, 2019.

[188] C. R. Glenn *et al.*, "Feasibility and acceptability of ecological momentary assessment with high-risk suicidal adolescents following acute psychiatric care," *Journal of Clinical Child & Adolescent Psychology*, vol. 51, no. 1, pp. 32–48, 2020.

[189] I. Gratch *et al.*, "Detecting suicidal thoughts: The power of ecological momentary assessment," *Depression and anxiety*, vol. 38, no. 1, pp. 8–16, 2020.

[190] M. A. Oquendo *et al.*, "Highly variable suicidal ideation: A phenotypic marker for stress induced suicide risk," *Molecular psychiatry*, vol. 26, no. 9, pp. 5079–5086, 2020.

[191] A. Porras-Segovia *et al.*, "Smartphone-based ecological momentary assessment (ema) in psychiatric patients and student controls: A real-world feasibility study," *Journal of Affective Disorders*, vol. 274, pp. 733–741, 2020.

[192] E. Singer and M. P. Couper, "Do incentives exert undue influence on survey participation? experimental evidence," *Journal of empirical research on human research ethics*, vol. 3, no. 3, pp. 49–56, 2008.

[193] S. W. Groth, "Honorarium or coercion: Use of incentives for participants in clinical research," *The Journal of the New York State Nurses' Association*, vol. 41, no. 1, p. 11, 2010.

[194] R. A. Bernert, C. L. Turvey, Y. Conwell, and T. E. Joiner, "Association of poor subjective sleep quality with risk for death by suicide during a 10-year period: A longitudinal, population-based study of late life," *JAMA psychiatry*, vol. 71, no. 10, pp. 1129–1137, 2014.

[195] S. X. Li *et al.*, "Sleep disturbances and suicide risk in an 8-year longitudinal study of schizophrenia-spectrum disorders," *Sleep*, vol. 39, no. 6, pp. 1275–1282, 2016.

[196] R. A. Bernert, M. A. Hom, N. G. Iwata, and T. E. Joiner, "Objectively assessed sleep variability as an acute warning sign of suicidal ideation in a longitudinal evaluation of young adults at high suicide risk," *The Journal of clinical psychiatry*, vol. 78, no. 6, p. 19 738, 2017.

[197] A. Mirsu-Paun, I. Jaussent, G. Komar, P. Courtet, and J. Lopez-Castroman, "Sleep complaints associated with wish to die after a suicide crisis—an exploratory study," *Journal of sleep research*, vol. 26, no. 6, pp. 726–731, 2017.

[198] D. Littlewood, S. Kyle, L. Carter, S. Peters, D. Pratt, and T. Gooding, "Short sleep duration and poor sleep quality predict next-day suicidal ideation: An ecological momentary assessment study," *Psychological Medicine*, vol. 49, no. 3, 2018.

[199] W. R. Pigeon, M. Pinquart, and K. Conner, "Meta-analysis of sleep disturbance and suicidal thoughts and behaviors," *The Journal of clinical psychiatry*, vol. 73, no. 9, p. 11 734, 2012.

[200] S. Malik *et al.*, "The association between sleep disturbances and suicidal behaviors in patients with psychiatric diagnoses: A systematic review and meta-analysis," *Systematic reviews*, vol. 3, no. 1, pp. 1–9, 2014.

[201] A. Porras-Segovia *et al.*, "Contribution of sleep deprivation to suicidal behaviour: A systematic review," *Sleep Medicine Reviews*, vol. 44, pp. 37–47, 2019.

[202] L. S. Chaïb, A. P. Segovia, E. Baca-Garcia, and J. Lopez-Castroman, "Ecological studies of sleep disturbances during suicidal crises," *Current psychiatry reports*, vol. 22, no. 7, pp. 1–8, 2020.

[203] H.-T. Lin *et al.*, "Insomnia as an independent predictor of suicide attempts: A nationwide population-based retrospective cohort study," *BMC psychiatry*, vol. 18, no. 1, pp. 1–11, 2018.

[204] T. Forkmann and T. Teismann, "Entrapment, perceived burdensomeness and thwarted belongingness as predictors of suicide ideation," *Psychiatry research*, vol. 257, pp. 84–86, 2017.

[205] E. Czyz, C. Glenn, D. Busby, and C. King, "Daily patterns in nonsuicidal self-injury and coping among recently hospitalized youth at risk for suicide," *Psychiatry research*, vol. 281, p. 112 588, 2019.

[206] E. Peters *et al.*, "Instability of suicidal ideation in patients hospitalized for depression: An exploratory study using smartphone ecological momentary assessment.," *Archives of Suicide Research: Official Journal of the International Academy for Suicide Research*, vol. 26, no. 1, pp. 56–69, 2020.

[207] Y. Kitagawa *et al.*, "Appetite loss as a potential predictor of suicidal ideation and self-harm in adolescents: A school-based study," *Appetite*, vol. 111, pp. 7–11, 2017.

[208] J. C. Franklin *et al.*, "Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research.," *Psychological bulletin*, vol. 143, no. 2, p. 187, 2017.

[209] W. M. Association *et al.*, "World medical association declaration of helsinki: Ethical principles for medical research involving human subjects," *Jama*, vol. 310, no. 20, pp. 2191–2194, 2013.

[210] K. Posner *et al.*, "The columbia–suicide severity rating scale: Initial validity and internal consistency findings from three multisite studies with adolescents and adults," *American journal of psychiatry*, vol. 168, no. 12, pp. 1266–1277, 2011.

[211] C. Fartacek, G. Schiepek, S. Kunrath, R. Fartacek, and M. Plöderl, "Real-time monitoring of non-linear suicidal dynamics: Methodology and a demonstrative case report," *Frontiers in psychology*, vol. 7, p. 130, 2016.

[212] C. H. Bastien, A. Vallières, and C. M. Morin, "Validation of the insomnia severity index as an outcome measure for insomnia research," *Sleep medicine*, vol. 2, no. 4, pp. 297–307, 2001.

[213] M.-M. G. Wilson *et al.*, "Appetite assessment: Simple appetite questionnaire predicts weight loss in community-dwelling adults and nursing home residents," *The American journal of clinical nutrition*, vol. 82, no. 5, pp. 1074–1081, 2005.

[214] D. V. Sheehan *et al.*, "The mini-international neuropsychiatric interview (mini): The development and validation of a structured diagnostic psychiatric interview for dsm-iv and icd-10," *Journal of clinical psychiatry*, vol. 59, no. 20, pp. 22–33, 1998.

[215] D. J. Buysse, C. F. Reynolds III, T. H. Monk, S. R. Berman, and D. J. Kupfer, "The pittsburgh sleep quality index: A new instrument for psychiatric practice and research," *Psychiatry research*, vol. 28, no. 2, pp. 193–213, 1989.

[216] A. J. Rush, C. M. Gullion, M. R. Basco, R. B. Jarrett, and M. H. Trivedi, "The inventory of depressive symptomatology (ids): Psychometric properties," *Psychological medicine*, vol. 26, no. 3, pp. 477–486, 1996.

[217] R. C. Young, J. T. Biggs, V. E. Ziegler, and D. A. Meyer, "A rating scale for mania: Reliability, validity and sensitivity," *The British journal of psychiatry*, vol. 133, no. 5, pp. 429–435, 1978.

[218] C. D. Spielberger, "State-trait anxiety inventory for adults," 1983.

[219] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[220] K. Russell, S. Rasmussen, and S. C. Hunter, "Insomnia and nightmares as markers of risk for suicidal ideation in young people: Investigating the role of defeat and entrapment," *Journal of clinical sleep medicine*, vol. 14, no. 5, pp. 775–784, 2018.

[221] E. F. Ward-Ciesielski, E. S. Winer, C. W. Drapeau, and M. R. Nadorff, "Examining components of emotion regulation in relation to sleep problems and suicide risk," *Journal of affective disorders*, vol. 241, pp. 41–48, 2018.

[222] É. Fortier-Brochu, S. Beaulieu-Bonneau, H. Ivers, and C. M. Morin, "Insomnia and daytime cognitive performance: A meta-analysis," *Sleep medicine reviews*, vol. 16, no. 1, pp. 83–94, 2012.

[223] A. J. Krause *et al.*, "The sleep-deprived human brain," *Nature Reviews Neuroscience*, vol. 18, no. 7, pp. 404–418, 2017.

[224] R. C. O'Connor and O. J. Kirtley, "The integrated motivational–volitional model of suicidal behaviour," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 373, no. 1754, p. 20 170 268, 2018.

[225] R. M. Ng, M. Di Simplicio, F. McManus, H. Kennerley, and E. A. Holmes, "'flash-forwards' and suicidal ideation: A prospective investigation of mental imagery, entrapment and defeat in a cohort from the hong kong mental morbidity survey," *Psychiatry Research*, vol. 246, pp. 453–460, 2016.

[226] S. S. Berg, P. S. Rosenau, and J. R. Prichard, "Sleep quality mediates the relationship between traumatic events, psychological distress, and suicidality in college undergraduates," *Journal of American College Health*, pp. 1–4, 2020.

[227] Y. S. Yang, G. W. Ryu, and M. Choi, "Methodological strategies for ecological momentary assessment to evaluate mood and stress in adult patients using mobile phones: Systematic review," *JMIR mHealth and uHealth*, vol. 7, no. 4, e11215, 2019.

[228] S. Saeb, E. G. Lattie, S. M. Schueller, K. P. Kording, and D. C. Mohr, "The relationship between mobile phone location sensor data and depressive symptom severity," *PeerJ*, vol. 4, e2537, 2016.

[229] J. Asselbergs, J. Ruwaard, M. Ejdys, N. Schrader, M. Sijbrandij, H. Riper, *et al.*, "Mobile phone-based unobtrusive ecological momentary assessment of day-to-day mood: An explorative study," *Journal of medical Internet research*, vol. 18, no. 3, e5505, 2016.

[230] J. Torous *et al.*, "Utilizing a personal smartphone custom app to assess the patient health questionnaire-9 (phq-9) depressive symptoms in patients with major depressive disorder," *JMIR mental health*, vol. 2, no. 1, e3889, 2015.

[231] L. J. Hauge, K. Stene-Larsen, T. K. Grimholt, C. Øien-Ødegaard, and A. Reneflot, "Use of primary health care services prior to suicide in the norwegian population 2006–2015," *BMC health services research*, vol. 18, no. 1, pp. 1–7, 2018.

[232] H. C. Schulberg, M. L. Bruce, P. W. Lee, J. W. Williams Jr, and A. J. Dietrich, "Preventing suicide in primary care patients: The primary care physician's role," *General hospital psychiatry*, vol. 26, no. 5, pp. 337–345, 2004.

[233] R. L. Wasserstein, A. L. Schirm, and N. A. Lazar, *Moving to a world beyond "p< 0.05"*, 2019.

[234] M. A. Reger, I. H. Stanley, and T. E. Joiner, "Suicide mortality and coronavirus disease 2019—a perfect storm?" *JAMA psychiatry*, vol. 77, no. 11, pp. 1093–1094, 2020.

[235] E. A. Holmes *et al.*, "Multidisciplinary research priorities for the covid-19 pandemic: A call for action for mental health science," *The Lancet Psychiatry*, vol. 7, no. 6, pp. 547–560, 2020.

[236] A. Porras-Segovia *et al.*, "Disturbed sleep as a clinical marker of wish to die: A smartphone monitoring study over three months of observation," *Journal of affective disorders*, vol. 286, pp. 330–337, 2021.

[237] N. C. Jacobson *et al.*, "Flattening the mental health curve: Covid-19 stay-at-home orders are associated with alterations in mental health search behavior in the united states," *JMIR mental health*, vol. 7, no. 6, e19347, 2020.

[238] C. A. Claassen *et al.*, "Effect of 11 september 2001 terrorist attacks in the usa on suicide in areas surrounding the crash sites," *The British Journal of Psychiatry*, vol. 196, no. 5, pp. 359–364, 2010.

[239] M. Osman and A. C. Parnell, "Effect of the first world war on suicide rates in ireland: An investigation of the 1864–1921 suicide trends," *BJPsych open*, vol. 1, no. 2, pp. 164–165, 2015.

[240] G. D. Batty *et al.*, "Psychosocial characteristics as potential predictors of suicide in adults: An overview of the evidence with new results from prospective cohort studies," *Translational Psychiatry*, vol. 8, no. 1, pp. 1–15, 2018.

[241] S. Liu *et al.*, "Online mental health services in china during the covid-19 outbreak," *The Lancet Psychiatry*, vol. 7, no. 4, e17–e18, 2020.

[242] R. E. Schapire, Y. Freund, P. Bartlett, W. S. Lee, *et al.*, "Boosting the margin: A new explanation for the effectiveness of voting methods," *The annals of statistics*, vol. 26, no. 5, pp. 1651–1686, 1998.

[243] Z. Borsos, A. Krause, and K. Y. Levy, "Online variance reduction for stochastic optimization," *arXiv preprint arXiv:1802.04715*, 2018.

[244] F. Salehi, E. Celis, and P. Thiran, "Stochastic optimization with bandit sampling," *arXiv preprint arXiv:1708.02544*, 2017.

[245] H. Namkoong and J. C. Duchi, "Variance-based regularization with convex objectives," arXiv:1610.02581, 2016.

[246] D. Csiba and P. Richtárik, "Importance sampling for minibatches," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 962–982, 2018.

[247] Y. Wang, A. Kucukelbir, and D. M. Blei, "Robust probabilistic modeling with bayesian data reweighting," *arXiv preprint arXiv:1606.03860*, 2016.

[248] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, Ieee, 2009, pp. 248–255.

[249] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[250] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS workshop on deep learning and unsupervised feature learning*, vol. 2011, 2011, p. 5.

[251] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Citeseer, Tech. Rep., 2009.

[252] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition.," in *In Proc. ICLR*, 2015.

[253] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," *arXiv preprint arXiv:1302.4389*, 2013.

APPENDIX A. ICD-10

ICD-10 code	Definition
F0	Organic, including symptomatic, mental disorders
F1	Mental and behavioral disorders due to psychoactive substance use
F2	Schizophrenia, schizotypal and delusional disorders
F3	Mood affective disorders
F4	Neurotic, stress-related and somatoform disorders
F5	Behavioral syndromes associated with physiological disturbances and physical factors
F6	Disorders of adult personality and behavior
F7	Mental retardation
F8	Disorders of psychological development
F9	Behavioral and emotional disorders with onset usually occurring in childhood and adolescence

APPENDIX B. CGI SCORES: SEVERITY OF ILLNESS

Considering your total clinical experience with this particular population, how mentally ill is the patient at this time?

0 = Not assessed

1 = Normal, not at all ill

2 = Borderline mentally ill

3 = Mildly ill

4 = Moderately ill

5 = Markedly ill

6 = Severely ill

7 = Among the most extremely ill patients

APPENDIX C. COMPLETE EMA QUESTIONNAIRE AND SCORING

Number	Area	Question	Minimum value	Maximum value	Scoring
1	Negative feelings	Today I feel psychological pain	No pain	Maximum pain	1-7
2	Negative feelings	I feel stressed out today (with pressure, overwhelmed)	No stress	Maximum stress	1-7
3	Negative feelings	Today I feel restless (agitated), with the need to keep moving	No restlessness	Maximum restlessness	1-7
4	Negative feelings	Today I feel full of hope[15]	No hope	Maximum hope	1-7
5	Negative feelings	Today I have felt hatred or anger towards myself	No hatred	Maximum hatred	1-7
6	Negative feelings	Today I have felt hatred or anger towards others	No hatred	Maximum hatred	1-7
7	Passive suicidal ideation	Today I feel the wish to live[15]	No wish to live	Maximum wish to live	1-7
8	Passive suicidal ideation	Today I feel the wish to die	No wish to die	Maximum wish to live	1-7
9	Negative feelings	Today I wished I had a trusted person to tell my personal issues	Not at all	Absolutely	1-7
10	Negative feelings	Today I felt like a stranger (out of place)	Not at all	Absolutely	1-7
11	Negative feelings	Today I had the impression that important people around me want to decide for me what I should think and do	Not at all	Absolutely	1-7
12	Negative feelings	Today I have wished to receive more recognition and love from others	Not at all	Absolutely	1-7
13	Negative feelings	Today I believe I have contributed to the well-being of my family/friends[15]	Not at all	Absolutely	1-7
14	Negative feelings	Today I believe I have contributed to the well-being of the people around me[15]	Not at all	Absolutely	1-7
15	Negative feelings	Today I felt disconnected from the rest of the people	Not at all	Absolutely	1-7
16	Sleep problems	Last night I had trouble getting to sleep	None	Very severe	0-4
17	Sleep problems	Last night I had trouble staying asleep	None	Very severe	0-4
18	Sleep problems	This morning I had trouble with premature awakening	None	Very severe	0-4
19	Sleep problems	Currently, others think that sleep problems affect my quality of life	Not at all	Absolutely	1-7
20	Sleep problems	Today when I woke up, I felt ...[15]	Very bad	Very good	1-7
21	Sleep problems	Last night the quality of my sleep was ...[15]	Very bad	Very good	1-7
22	Sleep problems	Today I am satisfied with my sleep[15]	Very unsatisfied	Very satisfied	1-7
23	Sleep problems	I am currently worried or stressed about my sleep problems	Not at all	Very much	0-4
24	Sleep problems	Currently my sleep problems are interfering with my daily activity	Not at all	Very much	0-4
25	Sleep problems	Today I feel tired during the day because of my sleep problems	Not at all	Absolutely	1-7
26	Appetite	In the last few days my appetite is ...[15]	Very little	Very big	0-4
27	Appetite	In the last days when I eat I feel full after eating ...[15]	Only a few bites	Almost never	0-4
28	Appetite	In the last few days I have been hungry[15]	Never	All the time	0-4
29	Appetite	In the last days when I eat the food tastes ...[15]	Very bad	Very good	0-4
30	Appetite	Compared to some years ago, nowadays food tastes ...[15]	Much worse	Much better	0-4
31	Appetite	In the last few days I usually do ...[15]	Less than one meal a day	More than three meals a day	0-4
32	Appetite	In the last few days when I eat, I feel sick or nauseous ...[15]	Most times	Never	0-4

[15]The scores from these items were reversed during data processing.